DOI: 10.1111/risa.14002

ORIGINAL ARTICLE

Online patient feedback as a safety valve: An automated language analysis of unnoticed and unresolved safety incidents

Alex Gillespie[1,2] | **Tom W. Reader**[1]

[1]Department of Psychological & Behavioural Science, London School of Economics, London, UK

[2]Department of Psychology, Oslo New University College, Oslo, Norway

Correspondence
Alex Gillespie, Department of Psychological & Behavioural Science, London School of Economics, Houghton Street, London, WC2A 2AE, UK.
Email: a.t.gillespie@lse.ac.uk

Abstract

Safety reporting systems are widely used in healthcare to identify risks to patient safety. But, their effectiveness is undermined if staff do not notice or report incidents. Patients, however, might observe and report these overlooked incidents because they experience the consequences, are highly motivated, and independent of the organization. Online patient feedback may be especially valuable because it is a channel of reporting that allows patients to report without fear of consequence (e.g., anonymously). Harnessing this potential is challenging because online feedback is unstructured and lacks demonstrable validity and added value. Accordingly, we developed an automated language analysis method for measuring the likelihood of patient-reported safety incidents in online patient feedback. Feedback from patients and families (n = 146,685, words = 22,191,427, years = 2013–2019) about acute NHS trusts (hospital conglomerates; n = 134) in England were analyzed. The automated measure had good precision (0.69) and excellent recall (0.98) in identifying incidents; was independent of staff-reported incidents ($r = -0.04$ to 0.19); and was associated with hospital-level mortality rates (z = 3.87; $p < 0.001$). The identified safety incidents were often reported as unnoticed (89%) or unresolved (21%), suggesting that patients use online platforms to give visibility to safety concerns they believe have been missed or ignored. Online stakeholder feedback is akin to a safety valve; being independent and unconstrained it provides an outlet for reporting safety issues that may have been unnoticed or unresolved within formal channels.

KEYWORDS
incident-reporting, natural language processing, online feedback, patient safety

1 | INTRODUCTION

Approximately 10% of hospital patients experience an adverse event during treatment (unintended harm due to errors), such as exacerbating resource pressure, harm, and even mortality (Lane et al., 2021; Makary & Daniel, 2016; National Academies of Sciences & Medicine, 2018; Vincent et al., 2001; World Health Organization, 2017). To reduce adverse events, healthcare organizations have invested in safety reporting systems for staff to report observations or involvement in safety incidents (adverse events and near misses) in order to identify and mitigate emerging risks (Barach & Small, 2000; Vincent et al., 2017). However, the success of these reporting systems in reducing adverse events has been limited due to inconsistencies in staff recognizing and reporting incidents (Shojania & Thomas, 2013; Stavropoulou et al., 2015).

Patient and family reports of care submitted to healthcare review websites (henceforth "online patient feedback") can augment risk management in hospitals (Greaves et al., 2013; Griffiths & Leaver, 2017). Specifically, we propose that online feedback is especially valuable for monitoring unnoticed and unresolved safety incidents. To this end, we introduce and validate an automated language analysis

methodology for identifying safety incidents in the narrative text of online patient feedback about treatment experiences in UK National Health Service (NHS) hospitals. To understand the unique value of these reports, we investigate the extent to which they report unnoticed and unresolved safety incidents. We also use this measure to investigate the relationship between patient-reported safety incidents online and staff-reported safety incidents, and we examine associations with hospital-level mortality rates. The results suggest that online patient feedback may act as a safety valve, capturing safety issues that patients perceive to be unnoticed or unresolved.

2 | LITERATURE REVIEW

The development of robust safety reporting systems is fundamental to risk management in safety critical industries (e.g., aviation, nuclear power; Reason, 1998). Typically, reporting systems rely on staff submitting accounts of incidents that either lead to harm or the potential for harm. These reports are analyzed to monitor risk, understand causes, and guide organizational learning (Leaver & Reader, 2016).

Safety reporting systems aim to support organizational learning, which refers to adaptation and change within institutions through social processes that facilitate knowledge sharing, innovation, and collective action (Rashman et al., 2009). Analysis of safety incidents can identify common types of safety problems (e.g., workplace injuries, flight control errors) and their proximal causes (e.g., ergonomic design, ineffective teamwork), and, thus, support managers specify and remedy common factors underlying unsafe events (Macrae, 2009; O'Connor et al., 2007).

However, organizational learning is challenging, rarely occurring due to the mere circulation of information (Argyris, 1990; Stanton et al., 2017). In healthcare, despite frequent in-depth investigations into the causes of incidents (Kellogg et al., 2017), organizational learning has been especially difficult (Sujan et al., 2017). To address this challenge, safety incident data can help at a meta-level by identifying problems in the process for correcting problems (Hald et al., 2021). Identifying unresolved and persistent risks can support identifying underlying causes (e.g., policy, culture) and thus guide organizational change (Bisbey et al., 2021; Catino & Patriotta, 2013). Thus, reporting systems can support double-loop learning, that is, helping organizations to learn from problems (Argyris & Schön, 1978; Bateson, 1972).

Even when organizational learning does not occur, data on safety incidents can be useful by providing leading rather than lagging indications of safety issues (Walker, 2017). Consistent with the theory on high reliability organizations and hazard analysis (Hulebak & Schlosser, 2002; Leveson et al., 2009; Weick & Sutcliffe, 2011), reporting systems can provide prospective data on emerging risks (e.g., near-miss rates, types, locations) and unresolved problems (e.g., concerns raised). Such timely incident data can support pre-emptive action to avoid problems cascading and more serious incidents (Billings, 1999; Walker, 2017).

While the effectiveness of safety reporting systems varies (e.g., due to staff motivation, institutional commitment, challenges converting information into learning), these systems are a key part of safety management within most safety critical industries (e.g., aviation; Jausan et al., 2017). Consequently, safety reporting systems have been instituted within most healthcare organizations, as part of the drive to improve patient safety by adapting the methodologies of high-reliability industries (Barach & Small, 2000).

2.1 | Safety reporting systems in healthcare

Healthcare organizations (e.g., hospitals, national providers, regulators) have developed safety reporting systems for staff to document incidents in which patients were or could have been harmed (Stavropoulou et al., 2015). This has generated vast amounts of data that can be used by clinicians and managers to detect risks to patient safety (e.g., for procedures or hospital units), develop safety interventions (e.g., equipment design, procedures), and encourage organizational learning (Benn et al., 2009; Frey et al., 2002; Herzer et al., 2012; Mitchell et al., 2016).

Yet, the contribution of safety reporting systems in reducing adverse events in hospitals has been limited (Shojania & Thomas, 2013). Despite their volume, incident data are not associated with hospital outcomes (e.g., excess mortality, retrospective care record reviews) and are an unreliable indicator of risk (Howell et al., 2015). Furthermore, and perhaps explaining this observation, safety reports often provide only partial accounts of incidents and near-misses within hospitals (e.g., in terms of causes and prevalence; Van Dael et al., 2021), and have failed to warn of systemic patient safety failures (Macrae, 2016; Papanicolas & Figueroa, 2019). Finally, while safety reporting systems have been used to develop local safety interventions (e.g., team processes; Howell et al., 2017), they have yet to lead to more substantial changes (e.g., culture change) or reductions in adverse event rates (Mitchell et al., 2016; Stavropoulou et al., 2015).

Three main factors limit the effectiveness of safety reporting systems in healthcare. First, some safety incidents during healthcare treatments are difficult for staff to notice and thus report on (Millman et al., 2011; Sari et al., 2007). These difficult-to-monitor "blindspots" include problems in patients accessing care, errors in clinical notes, errors in post-care planning, cascading errors that span the patient's journey, and miscommunication issues (Gillespie & Reader, 2018). Second, organizational culture can undermine the effectiveness of reporting systems. For example, defensiveness, low safety standards, and concealing errors due to fear of consequences can create skewed data (Dixon-Woods et al., 2009; Gillespie, 2020; Lawton & Parker, 2002; Waring, 2005). Finally, reporting behaviors are mediated by the efficacy of safety systems: where healthcare staff do not believe these systems are effective, prioritized, or drivers of change, they are less likely to report incidents, thus further contributing to unreliable

incident data (Benn et al., 2009; Mitchell et al., 2016; Pfeiffer et al., 2010).

The problems with safety reporting systems in healthcare have also been found in other settings (e.g., aviation, energy, mining; Jausan et al., 2017). However, a distinctive feature of healthcare is that safety incidents are experienced directly by patients and families. This proximity to error has led to the suggestion that healthcare organizations could supplement staff-reported incident data with information collected from patients and families about safety problems experienced during treatments (Armitage et al., 2018; Reader & Gillespie, 2021). This is consistent with the stakeholder theory, which proposes that organizational governance (e.g., on risk) can be improved through the input of public stakeholders, as these stakeholders experience an institution's successes and failures, have alternative insight into potential solutions (e.g., from the perspective of end-users), have different speaking-up constraints, and contribute independence and diversity of thought to decision-making processes (Beierle, 2002; Freeman, 1984). In particular, online platforms where patients and families can report on poor healthcare experiences may provide additional insight into safety performance within hospitals (Boylan, Turk et al., 2020; Griffiths & Leaver, 2017).

2.2 | Safety reporting in online feedback by patients and families

Online feedback platforms are used by patients and families to report on experiences of treatments within hospitals (Mazanderani et al., 2021). These platforms provide a forum for patients and families to report good and poor experiences, giving hospitals the opportunity to respond and learn (e.g., thanking compliments, acknowledging mistakes; Ramsey et al., 2019). Despite the challenges in extracting insights from these unstructured narratives (Zakkar & Lizotte, 2021), these data have the potential to provide rich patient-centered insights about hospital safety.

First, online patient feedback may report incidents that staff do not observe or report. Analyses of patients' formal complaints has found that they reveal safety issues difficult for staff to notice, such as, problems in accessing and exiting care, treatment omissions and neglect, miscommunication during diagnoses, incorrect patient notes, lapses in delivering medications, and continuity of care failures (Gillespie & Reader, 2018). Furthermore, because patients and families are independent of factors that shape staff reporting (e.g., normalization of risk-taking, organizational culture), they may report on incidents that staff overlook or are reluctant to report. Moreover, because online forums are independent and allow patients to remain anonymous (i.e., avoiding confronting staff directly) they may foster patient voices that are shy, hesitant, or intimidated. Therefore, online patient feedback may contribute to safety reporting systems through providing unconstrained, albeit noisy, data on previously invisible and "unnoticed" safety incidents.

Second, online platforms provide a "last resort" for patients and families to report on safety issues that they feel have not been addressed or taken seriously. Public inquiries into major healthcare safety failures repeatedly reveal poor culture, half-hearted investigations, and even suppression of complaints from both patients and staff (Kirkup, 2015; Ockenden, 2022). In this context, online platforms may capture issues that patients may be reluctant or dissuaded to report face-to-face or formally (e.g., due to fears of jeopardizing ongoing care; Doherty & Stavropoulou, 2012; Entwistle et al., 2010). Thus, open online platforms, that allow for anonymity (Locock et al., 2020), provide patients and families with a forum to make public incidents they believe have been dismissed or otherwise "unresolved" within hospitals.

In sum, we propose that online patient feedback can provide information on unnoticed and unresolved safety incidents, thereby potentially supporting healthcare organizations to develop more holistic and robust analyses of safety risks. We advance this idea by developing and testing a methodology for identifying safety incidents reported by patients and families in online feedback to NHS hospitals in England. More specifically we conceptualize and evidence how independent online feedback from stakeholders (i.e., patients) can add value by augmenting staff reporting systems and supporting risk management in organizations.

3 | CURRENT STUDY

Our first research question (RQ1) examined whether a valid methodology could be developed for identifying and analyzing safety incidents reported in online feedback at scale. The data source was the "Care Opinion" website, which has been used by hundreds of thousands of patients and families to report feedback on healthcare organizations in the UK NHS. Despite the richness of this narrative feedback, questions have been raised about its validity (especially for identifying safety incidents), accuracy, and representativeness, with the anonymity undermining its legitimacy for clinicians (Bjertnaes et al., 2020; Locock et al., 2020; Patel et al., 2015). A further issue is how to effectively analyze the high volumes of unstructured, open-ended textual data (e.g., from many thousands of patients, on a wide range of issues) submitted to online feedback portals in order to identify safety incidents (Boylan, Turk et al., 2020). To address these concerns, automated language analysis can be used to reliably distill "messy" patient data into clear and actionable insights (Giardina et al., 2018; Gibbons & Greaves, 2018; Griffiths & Leaver, 2017). This strategy involves using text search algorithms to identify words and sentences consistent with a given topic (e.g., safety incidents) within unstructured text. Such an approach has increasingly been used to identify safety problems reported by members of the public in online forums in diverse domains (Abrahams et al., 2012; Bleaney et al., 2018; Goldberg et al., 2020). Automated textual analysis can facilitate the identification and extraction of text (i.e., sentences, paragraphs, posts) relating to safety incidents within

online patient feedback. Specifically, algorithms can be used to search vast quantities of feedback for safety incidents, thus supporting risk monitoring by benchmarking organizations and supporting learning by surfacing the most safety-relevant feedback.

However, at present, there is no validated methodology for analyzing reports of safety incidents within online patient feedback (Marsh et al., 2019). Accordingly, we developed a text search algorithm to identify potential reports of safety incidents by patients and families within the Care Opinion data. Specifically, we utilized a word embedding approach, which uses models of the statistical relations between words in a language (see Section 4.2) to create an automated textual measure of patient-reported safety incidents that identifies sentences highly associated with terms relating to safety incidents in healthcare. To evaluate the validity of this measure, we identified examples of patient feedback that were high- or low-scoring in terms of our textual algorithm, and manually examined whether these were distinguished by their reporting of safety incidents experienced during healthcare treatment. Accordingly, RQ1 was: *Can an automated textual measure accurately identify patient-reported safety incidents in online feedback?*

Our second research question (RQ2) investigated whether the safety incidents reported by patients and families through online feedback were related to unnoticed or unresolved incidents. The key benefit of patient-reported incidents is that they might contain safety events that go unnoticed or unresolved by staff, thereby supplementing existing safety reporting systems (Giardina et al., 2018; Van Dael et al., 2021). We qualitatively analyzed the safety incidents identified using the automated textual measure in terms of whether they referred to safety incidents that were not reported by staff (unnoticed) or were left unaddressed and not taken seriously (unresolved). Such a finding would help to explain the added value of online feedback vis-à-vis other sources of safety data. Accordingly, RQ2 was: *Does the automated textual measure of patient-reported safety incidents identify incidents that are reported by patients and families to be unnoticed or unresolved?*

Our third research question (RQ3) examined the added value of using the automated textual measure of patient-reported safety incidents to evaluate risks to patient safety within hospitals. If patients' online feedback reports adverse events that are not reported by staff (i.e., unnoticed or unresolved), then, we reasoned, there should be a disconnect between these reports and staff-reported safety incidents at the level of the hospital. Although this suggestion has been evidenced at a qualitative and local level within healthcare systems (e.g., comparing staff and patient reports on unsafe events within a single hospital), it has not been shown across a healthcare system using online data (Giardina et al., 2018; Levtzion-Korach et al., 2010; Van Dael et al., 2021). Accordingly, we examined the association between the patient-reported safety incidents identified using the automated measure and the safety reporting rates in UK hospitals. We expected minimal correlation due to patients identifying different safety incidents, and staff reporting being skewed by factors such as organizational culture. Accordingly, RQ3 was: *Is the automated measure of patient-reported safety incidents independent of staff-reported safety incidents for hospitals?*

Finally, a lack of association between patient-reported safety incidents and staff-reported safety incidents could be due to incorrect perceptions by patients and family members or due to the online feedback being inconsistent and unreliable (Boylan, Williams et al., 2020; Locock et al., 2020; Turk et al., 2020). If online reports do reveal both the prevalence of safety issues experienced by patients and families, and the failure to detect and address safety incidents, we speculated that this should indicate the effectiveness of safety management in hospitals and thus be associated with hospital-level outcomes. To test this, we examined whether patient-reported safety incidents online predicted hospital-level mortality rate—a proxy measure of patient safety outcomes due to capturing excess deaths (Shwartz et al., 2011)—independently of staff-reported safety incidents. Establishing this would provide criterion validity, suggesting that online feedback provides a valid and independent source of safety data. Accordingly, RQ4 was: *Is the measure of patient-reported safety incidents associated with hospital-level mortality rates?*

4 | METHOD

The research design was a retrospective analysis of online patient feedback for all acute NHS trusts (hospital conglomerates) in England for the years 2013−2019. This feedback was analyzed quantitatively using an algorithm and a subset was analyzed qualitatively using a manual classification scheme.

4.1 | Data collection

We collected five datasets. *Patient feedback* was collected from the NHS and Care Opinion websites—the main feedback portals in England. These platforms share data, so feedback from both is available on each. Both platforms monitor IP addresses and manually review feedback to prevent abuses. Patient feedback was operationalized as feedback pertaining to any acute NHS trust in England submitted by patients (or their family and friends) to either website between 2013 and 2019. *Staff-reported safety incidents* were collected from the UK National Reporting and Learning System. We obtained the number and rate (per 1000 bed days) of all and severe (long-term harm or death) safety incidents. *The Summary Hospital-level Mortality Indicator* (SHMI) is an official calculation of the ratio between the actual number of patient deaths (in hospital or within 30 days of discharge) and the expected number of patient deaths based on patient characteristics (e.g., age, gender, primary diagnosis, secondary diagnosis, comorbidities). *Hospital spells* refers to the number of admission/discharge sequences in an

acute trust in a year, reported as part of the SHMI dataset. *Hospital case severity* refers to the clinical severity of the patients served in a hospital. It was calculated using the ratio of expected hospital mortality over the number of hospital spells (both part of the SHMI dataset). *Trust regions* refer to the four main geographic regions of England (London, South, North, Midlands), obtained from the NHS Organization Data Service.

For RQ1 and RQ2 we used feedback-level data. We combined the datasets so that each row was a separate item of feedback, and where feedback was directed at two or more hospitals, we duplicated the feedback, with a row for each hospital. Each item of feedback was scored separately, and the 1000 highest and lowest scoring items were used for the manual analyses.

For RQ3 and RQ4 we used aggregated trust-year-level feedback data. Aggregation was necessary because staff-reported safety incidents and hospital-level mortality rates are organization-level datapoints. Scoring each item of feedback separately and then aggregating the scores to the organization level, would have ignored the huge wordcount differences between feedback items. Accordingly, we aggregated all the text within each trust-year, and then ran the algorithm on the aggregated text, thus weighting each word in the feedback equally. To ensure robust volumes of textual data, we removed all hospital–year rows with less than 10,000 words of patient feedback (aggregated for the year).

The data sources, guidelines for manually classifying the data, and the manually classified data are available on GitHub (https://github.com/473x/safetyIncidents_analysis), and can be accessed via the Supporting Information.

4.2 | Measuring patient-reported safety incidents

We used text search algorithms to develop a measure of patient-reported safety incidents. Specifically, we drew on word embedding methods (Mikolov et al., 2013) to measure the use of words relating to severe safety incidents by patients and families in online feedback.

Word embedding methods measure the distribution and meaning of words used in textual data. They use artificial intelligence language models, created from unsupervised training on millions of documents, that encode the semantic similarity of hundreds of thousands of words. The similarity of words is calculated by statistically comparing the contexts in which words appear (i.e., similar words tend to be surrounded by the same words). This enables the creation of a *distributed dictionary representation* (DDR; Garten et al., 2018) of words associated with a concept of interest (e.g., "died" and "incapacitated" as words associated with severe safety incidents). Then, target documents (e.g., online patient reviews) are analyzed in terms of the degree to which the words used within them are similar to the DDR. Word embedding methods are a significant advancement over traditional word-counting methods because they do not limit analyses to the presence or absence of specific words, and instead examine the degree to which every word within a discourse is similar to a construct of interest (Boyd & Schwartz, 2021). They are not skewed by low-relevance, high-frequency terms, facilitate drilling down to textual specifics, produce normally distributed scores, and can be applied to any length of text (Garten et al., 2018).

For example, one can create a DDR of words associated with severe incidents (e.g., "died," "incapacitated," "maimed") and then measure the statistical likelihood of the words in the patient feedback co-occurring with the target terms (see Figure 1 for an illustration). The DDR method can score sentences, paragraphs, whole items of feedback, or aggregations of feedback (e.g., at unit, hospital, or even regional levels) in terms of how close they are to the target DDR, with the final score being proportional to the wordcount. This method is particularly suited to analyzing online feedback because patients and families use diverse language that changes over time and context, making it difficult to construct exhaustive and accurate word lists. The DDR method focuses on semantic similarity, and thus is more robust to changing word meanings. This method has increasingly been used in healthcare (Wang et al., 2018), for example, to analyze electronic health records (Blanco et al., 2020), translate medical terminology into lay terminology (Gu et al., 2019), and improve measures of sentiment in online health-related forums (Carrillo-de-Albornoz et al., 2018). The limitations are that the method is relatively novel, and DDR measures are shaped by both the chosen language model and the target terms.

4.3 | Developing an automated measure of patient-reported safety incidents

To develop the automated measure of patient-reported safety incidents, we used GloVe (Global Vectors for word representation) embeddings because they are based on the co-occurrence of words (rather than, for example, trying to predict missing words) and have performed favorably on similarity tasks (Pennington et al., 2014). Specifically, we used a publicly available language model created from the Common Crawl dataset of webpages, comprising 684,831 unique vectors represented in 300 dimensions (Explosion, 2020). While word embeddings trained on medical texts perform marginally better than word embeddings trained on more general texts (e.g., Wikipedia, news, webpages) for biomedical tasks (Wang et al., 2018), we chose to use the latter because our data were made up of online posts by laypeople. Similarity was calculated as the cosine distance between the average vector for the target terms and the average vector for the target text, yielding a score ranging from 0 (least similar) to 1 (most similar).

We used a four-step methodology to create the target terms underlying the DDR measure. First, we identified key terms based on the patient safety literature relating to adverse events and how patients and families describe these (King

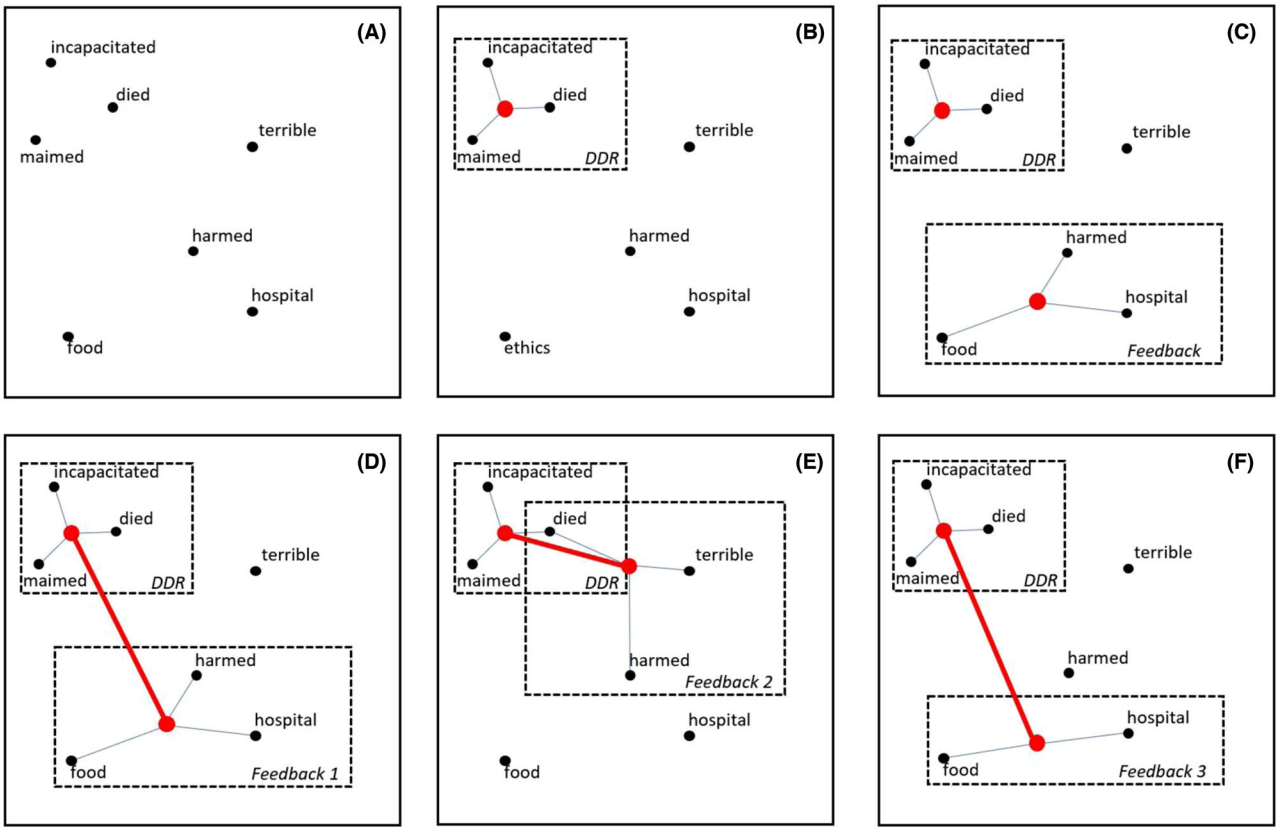

FIGURE 1 Illustration of using a DDR for severe safety incidents to score feedback. (A) illustrates a simplified language model (with seven words plotted in two dimensions; normally there would be hundreds of thousands of words in hundreds of dimensions). The distance between words encodes their semantic similarity. For example, "died" and "incapacitated" are close, but both are far away from "food." In (B), we define our DDR measure: a cluster of semantically related target terms (e.g., "died," "maimed," "incapacitated") selected to reliably and validly capture our target concept (i.e., severe incidents). The center-point of the DDR is the reference point in our measure. (C) introduces feedback 1 (containing three words "harmed," "hospital," and "food"). In order to measure the similarity of feedback 1 to the target terms, we locate the center-point of the words in the feedback. (D) illustrates measuring the distance from the center-point of the DDR to the center-point of feedback 1. (E) illustrates how the feedback with the words "died," "terrible," and "harmed" is more similar to the target terms for severe incidents. (F) illustrates how the feedback containing the words "hospital" and "food" is measured as less similar

et al., 2010; NHS Improvement, 2018). Second, we used the language model to generate similar candidate terms (e.g., "deceased" as similar to "died"). Third, we systematically examined high-scoring sentences for each target term and removed terms with low validity. For example, we did not include the word "accident" because high-scoring sentences often referred to the cause of a patient going to hospital and not a safety incident (e.g., a wrong medication). Finally, we scored the text using each target term separately, calculated inter-term reliability, and removed terms with low reliability (e.g., "safety", which seemed to be used to describe high-quality care rather than incidents).

The final list of target terms was: dead, deceased, died, dieing, disfigured, dying, grave, harmed, incapacitated, killed, killing, maimed, misdiagnosing, mortally, murdering, murderous, mutilated, overmedicated, perished, readmitted, scarred, succumb, succumbed, and traumatised. The terms "misdiagnosed," "readmitted," and "overmedicated" were included, despite having lower validity and reliability, in order to preserve the breadth of the measure. The list focuses on death-related terms because these identify the extremity of safety incidents (NHS Improvement, 2018). Less severe terms are not included because the aim of the DDR is to depict only the extremity (i.e., the extreme end of the measure). Less severe terms are captured by default because they will be similar to the extreme, but with a lower weighting (Garten et al., 2018). Some of the final terms are very strong (e.g., "killed," "murdering") and contain spelling mistakes (e.g., "dieing") because this is indicative of the sometimes-extreme language used by patients online.

4.4 | Data analysis

To assess the validity of the algorithm (RQ1), we manually classified the highest ($n = 1000$) and lowest ($n = 1000$) scoring feedback with over 100 words. Blinded to the algorithm's scores, the first author and a research assistant classified each item of feedback for the presence/absence of safety incidents, including identifying severe safety incidents (i.e.,

resulting in long-term harm or death). Interrater reliability for the manual classification was based on a sample of 100. The validity of the algorithm was assessed with a confusion matrix of scores (high/low), and manually categorized safety incidents (yes/no) were used to calculate precision and recall statistics.

To investigate whether the patient feedback was reporting unnoticed and unresolved incidents (RQ2), we manually analyzed the highest ($n = 1000$) and lowest ($n = 1000$) scoring feedback. The first author and a research assistant classified each individual item of feedback in terms of: (1) reporting an unnoticed safety incident if it included a safety incident and no mention of a hospital- or staff-initiated investigation; (2) reporting an unresolved safety incident if it included a safety incident and either mentioned that the issue had been dismissed by staff or directly asked safety-related questions that should have been addressed by staff (e.g., questioning a medication). Interrater reliability for the manual classification was based on a sample of 100. Descriptive statistics and qualitative analysis explored potentially unnoticed and unresolved safety incidents.

To examine associations between patient- and staff-reported safety incidents (RQ3), we calculated Spearman's rank correlations in the context of related variables, separately for each year.

To model the association between patient-reported safety incidents and hospital-level mortality rates (RQ4), we used a mixed linear model. The data were grouped by year and geographical region, with controls for hospital size (number of patient spells), feedback wordcount, hospital casemix severity, and feedback sentiment (Table 1).

The data, the Python code for the analysis, and a demonstration of the algorithm are available on GitHub (https://github.com/473x/safetyIncidents_analysis), and can be accessed via the Supporting Information.

5 | RESULTS

There were 146,685 individual items of feedback (22,191,427 words) pertaining to 134 acute English NHS trusts (out of 139 organizations in total) over 7 years. For RQ3 and RQ4 these were aggregated into 792 trust-year datapoints (each with at least 10 thousand words of feedback) and paired with the secondary data (see Table 2 for descriptive).

Figure 2 reports the hospital-level distributions staff-reported safety incident rates (all and severe, rate per 1000 bed days) and the patient-reported safety incident measure (similarity of patient feedback to terms indicating a safety incident). This reveals that staff-reported safety incident data (all and especially severe) were strongly skewed towards low rates, with little variance in low scores, and a spike of incidents in 2013. In contrast, the patient-reported safety incident measure is more consistent, with a non-skewed distribution, and good variance across the range of scores.

5.1 | Validity of the patient-reported safety incident measure (RQ1)

The interrater reliability for the manual classification was strong for all safety incidents ($\alpha = 0.89$, 95% CI = 0.84, 0.93) and moderate for severe safety incidents ($\alpha = 0.72$, 95% CI = 0.58, 0.81). There were 701 safety incidents (high-scoring = 685, low-scoring = 16), 180 of which were severe (high-scoring = 180, low-scoring = 0).

Comparing the manual classifications to the highest and lowest scoring feedback revealed good precision (0.69) and excellent recall (0.98) for all safety incidents, but weak precision (0.18) and recall (0.31) for severe safety incidents. Manual classification found 1.6% ($n = 16$) safety incidents and no severe safety incidents in the low-scoring feedback ($n = 1000$). Manual classification found 68.5% ($n = 685$) safety incidents and 18% ($n = 180$) severe safety incidents in the high-scoring feedback ($n = 1000$). Examples of safety incidents that patients reported leading to harm included medical machinery repeatedly breaking down (e.g., infusion pumps); failures in infection control (leading to MRSA); errors in medication safety; neglect (e.g., not treating wounds, patients left in unsanitary conditions); diagnostic errors (e.g., for cancer, pre-eclampsia, physical ailments); failures to put in place care plans; unsuccessful routine procedures (e.g., hip operations); test results incorrect, lost, or not being communicated to other hospital units (e.g., for neurological illnesses); crisis symptoms not being recognized (e.g., heart attack, strokes); injuries in hospital (e.g., falls); extreme delays (e.g., for urgent CT scans); and patients being discharged prematurely or without risk assessment (e.g., with late-stage dementia).

5.2 | Unnoticed and unresolved safety incidents (RQ2)

The interrater reliability for the manual classification was excellent for reports of safety investigations ($\alpha = 0.92$, 95% CI = 0.88, 0.95) and whether they were patient-initiated ($\alpha = 0.92$, 95% CI = 0.88, 0.95), and moderate for reports of safety issues being dismissed ($\alpha = 0.74$, 95% CI = 0.62, 0.83) and for feedback directly asking safety questions ($\alpha = 0.63$, 95% CI = 0.45, 0.75).

Most high-scoring feedback did not mention an official safety investigation (89%, $n = 889$); when mentioned, it tended to be patient-initiated (e.g., writing a complaint). Qualitative examination identified many incidents that had occurred in staff blindspots, thus making them difficult for staff to report. These included issues arising before admission (e.g., unable to access services) or after discharge (e.g., premature discharge resulting in readmission), cascading low-level problems (e.g., poor hygiene practices), and failures to coordinate between units and visits. Moreover, patients reported actions that could make incidents difficult for staff to identify, for example, getting a second opinion, choosing

TABLE 1 Research variables used in the mixed linear model

Variable	Description	Rationale
Year	From 2013 to 2019 inclusive.	Random effect, controlling for variations in the data across time because language use (and thus measures from distributed language models) vary in time (Rodman, 2020).
Region	London, South, North, and Midlands.	Random effect, controlling for variations in geographic region because language use may vary by region.
Hospital spells	A hospital spell begins when a patient is admitted and ends when he or she is discharged.	Fixed effect, controlling for the size of hospitals because hospital size is associated with hospital-level mortality rates (Reader & Gillespie, 2021).
Hospital case severity	Ratio of the expected number of hospital deaths over the number of hospital spells.	Fixed effect, controlling for the severity of the cases that the hospital treats.
Staff-reported incidents (all)	Staff are encouraged to report all safety incidents, regardless of the harm; reported as a rate per thousand provider spells.	Testing whether the patient-reported incidents merely duplicate the staff-based data.
Staff-reported incidents (severe)	Staff are mandated to report severe safety incidents; reported as a rate per thousand provider spells.	Testing whether the patient-reported incidents merely duplicate the staff-based data.
Feedback wordcount	Number of words in all patient feedback for a given trust for a given year.	Fixed effect, controlling for the volume of textual feedback because language analysis can be sensitive to the volume of textual data. (Krippendorff, 2019)
Feedback sentiment	The positive-to-negative emotion (from +1 to −1) of feedback for a given trust for a given year, scored using the VADER (Hutto & Gilbert, 2014) algorithm.	Fixed effect, controlling for the positive/negative sentiment of the feedback. Although it is difficult to separate safety incidents from sentiment (assuming more severe incidents are reported with more negative sentiment), we included this control to be conservative with the safety incident measure and ensure that it is not merely detecting sentiment rather than safety incidents per se.
Patient-reported safety incidents	A continuous measure of the similarity of patient narratives to words indicative of a safety incident (e.g., "died," "overmedicated," "misdiagnosed"), with scores between 0 (least similar) and 1 (most similar).	Independent variable: it is used to test whether the likelihood of patient feedback reported online containing a safety incident is associated with hospital-level mortality.
Summary Hospital Mortality Indicator (SHMI)	The ratio between the actual and expected number of patient deaths, with a ratio above 1 indicating more deaths than expected given the patient demographic.	Outcome variable: hospital-level mortality rates are a frequently used measure of overall hospital safety (Toffolutti & Stuckler, 2019).

to be readmitted to a different hospital, or transitioning to private healthcare.

Unresolved safety incidents were indicated by reports of staff ignoring patients' concerns in 21% ($n = 213$) of the high-scoring feedback. Moreover, 6% ($n = 61$) of high-scoring feedback directly asked safety questions that should have been addressed by staff (e.g., "why did my daughter die?", "why weren't we told our daughter had sepsis?"), indicating that these patients' concerns had been raised but remained unaddressed. Unresolved safety issues ranged from having concerns dismissed in face-to-face interactions (e.g., "when I told the nurse […] other nurses started laughing") to obstructions and delays when questions were raised (e.g., "yet again, we have had no response").

Qualitative analysis of the feedback revealed that several patients were posting online as a last resort. One theme was that a formal complaint had stalled (e.g., "I have put a formal complaint […] 6months later still had no response to questions I've asked") or there was dissatisfaction with the outcome of a complaint investigation (e.g., "my partner and I know what happened that day, and so do the midwife and the doctor but we haven't been able to get an acceptance of

TABLE 2 Descriptives for the trust–year data ($n = 792$)

Descriptives (per trust–year)	Mean	St. Dev.	Min	Max	Skew	Kurtosis
Hospital spells	70700.8	30203.83	17335	275055	1.601	5.46
Hospital case severity	0.033	0.006	0.015	0.05	−0.155	0.49
Staff-reported incidents (all, n)	10321.58	5092.28	1832	43733	1.61	4.42
Staff-reported incidents (all, rate)	47.33	16.95	15.5	155.73	1.73	4.45
Staff-reported incidents (severe, n)	42.33	32.04	1	215	1.99	5.31
Staff-reported incidents (severe, rate)	0.22	0.21	.004	1.78	3.24	14.44
Feedback wordcount	28019.48	14754.41	10980	97067	1.56	2.82
Feedback sentiment (score −1 to +1)	0.17	0.06	−0.02	0.34	−0.1	0.11
Patient-reported incident measure (0-1)*	0.34	0.004	0.33	0.35	0.12	0.08
Hospital-level mortality rate	1	0.1	0.62	1.24	−0.75	1.01

*The patient-reported incident measure is a continuous textual measure of the likelihood of a safety incident being reported in the patient feedback. It is measured by computing the semantic similarity of all the textual feedback for a trust–year to the safety incident target terms. A score of 0 is complete dissimilarity and a score of 1 is complete similarity (both extremes are impossibly rare).

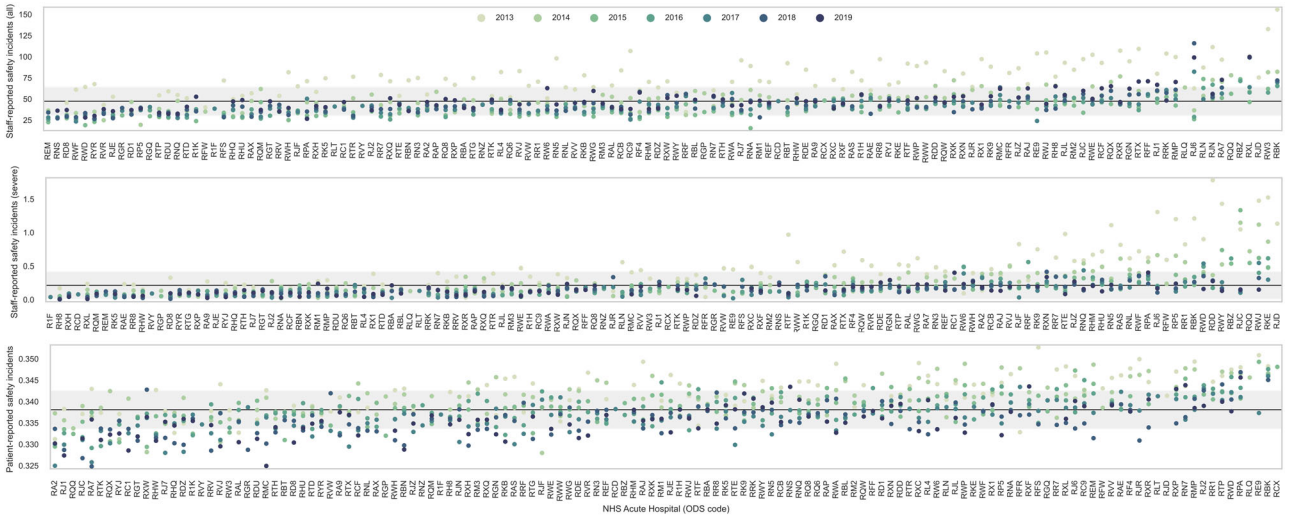

FIGURE 2 Staff-reported all (top, per 1000 bed days), staff-reported severe (middle, per 1000 bed days), and patient-reported (bottom, feedback similarity to target terms) incidents by trust. The shading indicates one standard deviation above and below the mean (the solid line is the mean, and trusts are ordered from lowest to highest mean score)

the truth"). Some patients mentioned that they had exhausted alternative avenues for having their concerns addressed (e.g., "when people like me post on here we are already worn down by having told managers and front-line workers what the problems are and have been dismissed").

5.3 | Correlations between the safety incident measures (RQ3)

To assess correlations between the safety incident measures, we calculated bivariate Spearman's rank correlations between all the variables for each year (Figure 3). Hospital spells were consistently associated with increased volumes of feedback (r range = 0.42, −0.68) and inconsistently associated with decreased sentiment (r range = −0.26, −0.1). Hospital case severity was associated with positive feedback sentiment (r range = 0.09, −0.4) and higher hospital-level mortality (r range = 0.17, −0.26). Feedback sentiment was consistently associated with reduced patient-reported incidents (r range = −0.57, −0.23). Patient-reported incidents were almost consistently associated with increased hospital-level mortality (r range = 0.14, 0.3). There was no significant association between patient- and staff-reported incidents (all r range = −0.04, 0.19; severe r range = −0.13, 0.14).

5.4 | Predicting hospital-level mortality (RQ4)

To test whether staff-reported and patient-reported safety incidents were predictive of hospital-level mortality, we

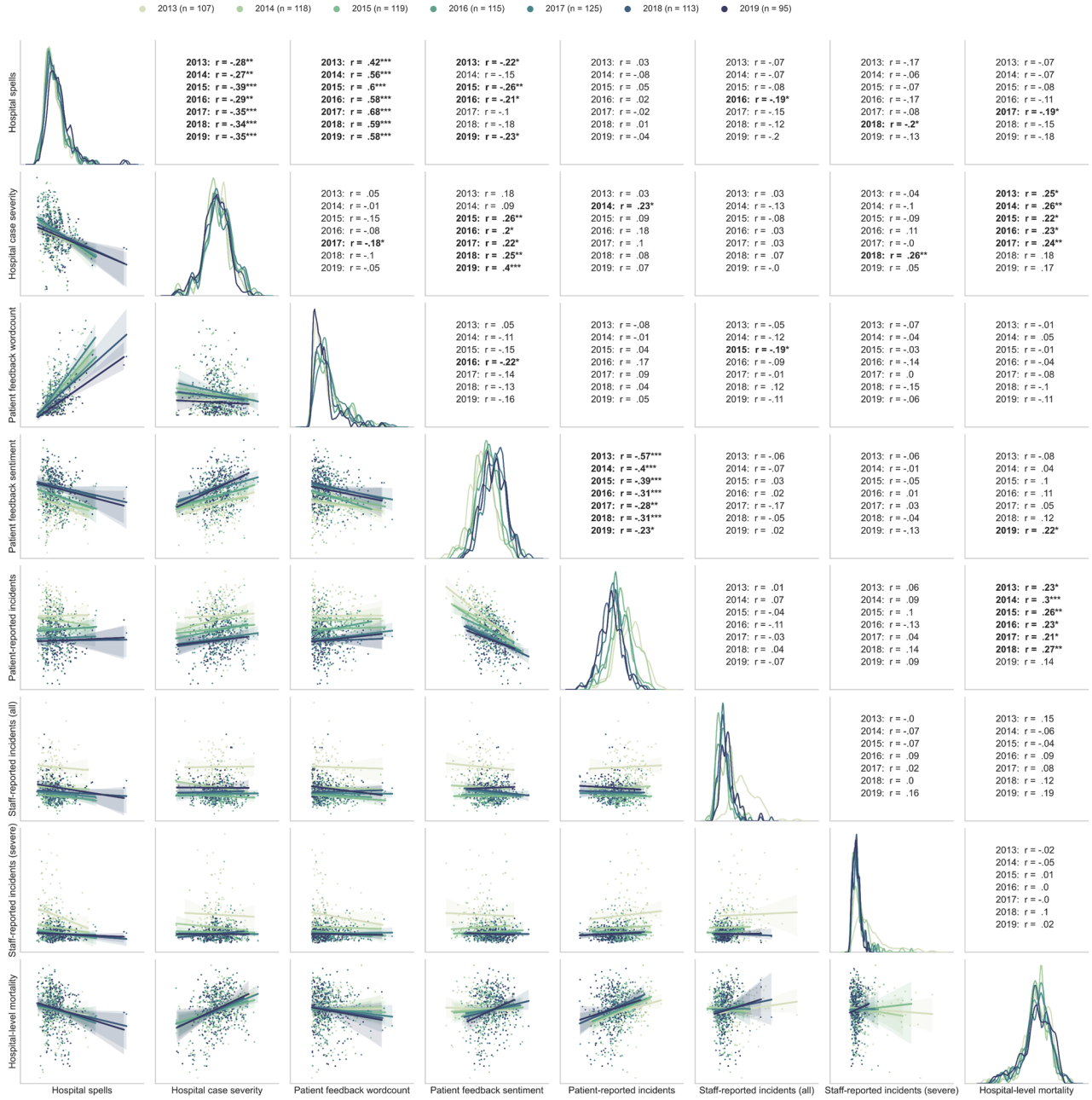

FIGURE 3 Spearman's rank correlations grouped by year (n is number of hospitals). The diagonal displays the distribution of each variable. Above the diagonal are the correlations (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$). Below the diagonal are scatterplots, with regression lines (95% CI) for each year

created a linear mixed effects model. The model was constructed using year and geographic region as crossed random effects, with hospital spells, hospital case severity, word count in the patient feedback, and a text-based measure of feedback sentiment as fixed effects.

Neither staff-reported measure of safety incidents (all, severe) were significantly predictive of hospital-level mortality. An increase of .1 in the all-incident measure was associated with an increase of .002 in hospital-level mortality (95% CI = −0.05, −0.05; $z = -0.07$; $p = 0.95$), holding all other variables fixed. Similarly, an increase of .1 in the severe-incident measure was associated with a decrease of 0.01 in hospital-level mortality (95% CI = −0.05, 0.03; $z = -0.57$; $p = 0.57$), holding all other variables fixed.

The measure of patient-reported safety incidents was found to be significantly predictive of hospital-level mortality (Figure 4). An increase of .1 in this measure was associated with an increase of .08in hospital-level mortality (95% CI = 0.04, 0.12; $z = 3.6$; $p < 0.001$). An increase of .01 in hospital-level mortality translates to an increase of 1% in actual deaths relative to the expected number of deaths based on patient characteristics.

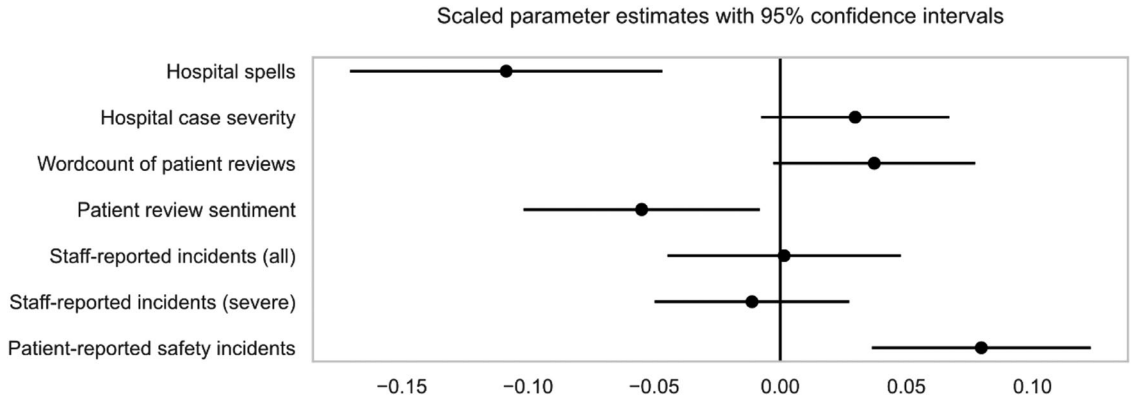

FIGURE 4 Scaled parameter estimates, with 95% confidence intervals, predicting hospital-level mortality rates

5.5 | Robustness checks

To assess whether the findings could be an artefact of the natural language processing method, we conducted three robustness checks. First, to assess the extent to which the language used to describe severe safety incidents varied across time, we calculated the mean test–retest reliability for year-pairs, which was comparable to previously recorded rates of transient error (McKenny et al., 2018). Second, to assess the extent to which target term selection shaped our overall findings, we reran the analysis with a random selection of half the target terms and with an expanded set of target terms; in both cases, the results were comparable. Third, given that word embeddings trained on context-relevant (i.e., medical) data have been shown to have greater accuracy than the generic GloVe model (Wang et al., 2018) that we used, we created a new language model based on our own data; this also produced the same pattern of findings. These checks are discussed in more detail in the Supporting Information.

6 | DISCUSSION

The results show that online patient feedback can provide information on unnoticed and unresolved safety incidents within hospitals, with these data being independent of staff-reported incidents and predictive of patient safety outcomes. Moreover, these patient-reported incidents can be extracted reliably and validly using automated algorithms. Using such algorithms can aid hospitals to develop more holistic and robust analyses of emerging and current risks to patient safety. At a theoretical level, the results suggest online patient feedback may function as a safety valve. Where patients experience unsafe treatments and believe that these have gone either unnoticed or unresolved, they report incidents online so that their concerns can be made visible.

Notably, the automated measure of patient-reported safety incidents was associated with hospital mortality, whereas staff-reported incidents were not. The latter lack of association may reflect the prior finding that the reliability of staff reporting systems is shaped by safety culture rather than the population of incidents (Howell et al., 2015; Jausan et al., 2017; Pfeiffer et al., 2010).

Our finding that patient-reported safety incidents are associated with hospital level mortality is consistent with prior research that has shown that online patient feedback is associated with healthcare outcomes (Placona & Rathert, 2021). However, our findings extend this literature by first sharpening the focus onto safety incidents and second identifying two complementary explanatory factors. First, online patient feedback may reliably track the number of safety incidents within a hospital by providing accurate information on adverse events. Second, online patient feedback may reveal hospitals that are poor at detecting and responding to safety incidents. In such cases, patients go online because they believe a hospital is not effective at addressing and recognizing safety problems (Boylan, Williams et al., 2020). Over a quarter of high-scoring feedback was related to patients and families reporting safety concerns that, from their perspective, had been dismissed, or asking safety questions that should have been addressed by hospital staff. Accordingly, in addition to tracking safety incidents per se, online feedback may also track organizations that fail to address and learn from safety incidents, which, in turn, leads to higher hospital-level mortality rates.

6.1 | Theoretical and practical implications

The findings suggest that online patient feedback could be valuable for supporting risk management in hospitals through providing alternative and supplementary information on safety incidents. Linking back to stakeholder theory (Freeman, 1984; Reader & Gillespie, 2021), patients and families—as the recipients of treatments and independent of hospitals—can bring diversity of understanding to patient safety. Their feedback can support healthcare staff, managers, and regulators to develop more holistic analyses of current and emerging risks. Neither staff nor patients have complete information, each group has unique insights and blindspots (Gillespie & Reader, 2018), and thus combining their perspectives is beneficial. Augmenting staff reports with patient

reports may be especially valuable when there is uncertainty about staff reporting. In such contexts patients reporting freely online (e.g., anonymously, without consequence) may act as a safety valve, revealing safety incidents that have been unnoticed or unresolved.

Monitoring patient-reported incidents online can make three contributions. First, it could provide reliable and valid data on safety incidents and thus enhance safety monitoring (Benn et al., 2009; Billings, 1999). This would support safety management by identifying areas for improvement not captured by staff reporting (e.g., relating to patient journeys) and identifying units where there is a heightened risk to patients. Second, it could support organizations' double-loop learning about learning from incidents. Providing information about problems does not inevitably lead to learning (Argyris, 1990; Stanton et al., 2017), especially in healthcare (Sujan et al., 2017). Identifying unnoticed and unresolved safety incidents, and discrepancies between staff and patient reports, might reveal problems in reporting practices (e.g., incidents not being logged) or failures in learning from mistakes. Third, it could be used to trigger preventative interventions. Given the challenge of turning lagging information into learning (Stanton et al., 2017), it is important to identify leading indicators (Walker, 2017). Many of the unnoticed and unresolved issues reported the online feedback were ongoing (e.g., treatment delays, dismissed symptoms). Thus, timely identification of these incidents could prevent them cascading towards more severe adverse events.

Natural language processing can support these contributions. Despite the challenges to analyzing unstructured narrative feedback (Boylan, Williams et al., 2020; Zakkar & Lizotte, 2021), it is possible to automatically identify patient-reported safety incidents reliably and validly. These algorithms can boost the rigor, efficiency, and scale of analyses, while preserving the ability to drill down into textual specifics (e.g., for qualitative analysis). They can be used to track the persistence of problems (i.e., failures to learn); provide near real-time analysis of feedback to preemptively identify unnoticed or unresolved incidents before they cascade towards more severe consequences; and make patterns in patient-reported safety incidents visible (e.g., to the broader public and regulators) thus motivating hospitals to address recuring problems.

Beyond healthcare, the observation that the service users of an organization can report on unnoticed and unresolved safety problems is significant. Research in several domains (e.g., food safety, product safety, transport, building safety, policing) has recognized that, like healthcare, public stakeholders may possess important insight into safety within organizations (Abrahams et al., 2012; Bleaney et al., 2018; Goldberg et al., 2020). Indeed, Turner's (1976) seminal theorization of accident development describes, through case studies, how the raising of safety concerns by members of the public—and the subsequent rejection by organizations—is a common feature of accidents (e.g., dismissing complaints). This finding has been borne out by numerous accident investigations and reinforces the insight that public stakeholders often possess important safety information (Hald et al., 2021). For example, public stakeholders can support risk management in consumer products (e.g., baby toys, vehicles; Abrahams et al., 2012; Bleaney et al., 2018), policing (e.g., violence; Dugan & Breda, 1991), transport (e.g., unsafe driving; Öz et al., 2014), hospitality (e.g., unhygienic food; Goldberg et al., 2020), and building safety (e.g., identifying fire risks; Cornish, 2021; MacLeod, 2018). The growth of online portals for public stakeholders to provide feedback, including on safety and risk, may represent a route through which to improve risk management. Specifically, online stakeholder feedback may, if analyzed carefully, provide insight on safety issues that have been filtered, suppressed, or otherwise inhibited in other channels of feedback. The unfiltered, unconstrained, and often anonymous nature of online stakeholder feedback is simultaneously a challenge for analysis but also central to its added value.

6.2 | Limitations

Patient-reported data are not neutral and are shaped by regional culture, publicity about a hospital, staff behavior (e.g., not responding to a complaint), and ideological commitments (e.g., for or against a national health service). We could not verify the safety incidents reported or whether they were genuinely unnoticed or unresolved. Our measure of patient-reported safety incidents is continuous, not categorical like staff-reported incidents. We did not have access to the text of the staff-reported incidents, so could not examine the text relating to the incidents being reported (i.e., to directly compare with patient online feedback). Finally, SHMI is only one limited criterion for safety (Lilford et al., 2004), and future studies may use alternative measures (e.g., patient case record reviews).

6.3 | Conclusion

Online patient feedback can supplement safety reporting systems in healthcare through providing information on safety incidents that go unnoticed and unresolved in hospitals. Medical error remains a leading cause of death in hospital (Makary & Daniel, 2016). Although safety reporting systems have led to local learning, they do not reveal incident prevalence in organizations (Howell et al., 2017) and have failed to provide early warnings of systemic failings. A key issue is that staff reporting can be unreliable (e.g., mediated by the local safety culture). Reports on safety incidents within online patient feedback provide an alternative and independent channel for monitoring risk. We found that natural language processing can reliably identify the likelihood of online patient feedback reporting a safety incident and validly predict hospital safety outcomes (independent of safety incidents). This is significant for the broader risk literature, as public stakeholders are increasingly raising concerns about safety through online channels (e.g., as customers or

service-users). We propose that the sharpest value of online feedback is that it acts as a safety valve for stakeholders to report safety incidents that they feel have gone either unnoticed or unresolved, thereby providing a unique additional line of defense in detecting and monitoring risk.

ACKNOWLEDGMENTS

The research was not funded by an external body. We would like to thank Care Opinion for facilitating access to the patient feedback data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Alex Gillespie https://orcid.org/0000-0002-0162-1269
Tom W. Reader https://orcid.org/0000-0002-3318-6388

REFERENCES

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, *54*(1), 87–97. https://doi.org/10.1016/j.dss.2012.04.005

Argyris, C. (1990). *Overcoming organizational defenses*. Allyn and Bacon.

Argyris, C., & Schön, D. A. (1978). *Organizational learning: A theory of action perspective*. Addison-Wesley.

Armitage, G., Moore, S., Reynolds, C., Laloë, P.-A., Coulson, C., McEachan, R., Lawton, R., Watt, I., Wright, J., & O'Hara, J. (2018). Patient-reported safety incidents as a new source of patient safety data: An exploratory comparative study in an acute hospital in England. *Journal of Health Services Research & Policy*, *23*(1), 36–43. https://doi.org/10.1177/2F1355819617727563

Barach, P., & Small, S. D. (2000). Reporting and preventing medical mishaps: Lessons from non-medical near miss reporting systems. *BMJ: British Medical Journal*, *320*(7237), 759. https://doi.org/10.1136/bmj.320.7237.759

Bateson, G. (1972). *Steps to an ecology of mind*. Ballantine Books.

Beierle, T. C. (2002). The quality of stakeholder-based decisions. *Risk Analysis: An International Journal*, *22*(4), 739–749. https://doi.org/10.1111/0272–4332.00065

Benn, J., Koutantji, M., Wallace, L., Spurgeon, P., Rejman, M., Healey, A., & Vincent, C. (2009). Feedback from incident reporting: Information and action to improve patient safety. *BMJ Quality & Safety*, *18*(1), 11–21. https://doi.org/10.1136/qshc.2007.024166

Billings, C. E. (1999). The NASA aviation safety reporting system: Lessons learned from voluntary incident reporting. In Scheffler, A., & Zipperer, L. A. (Eds.). *Enhancing patient safety and reducing errors*. National Patient Safety Foundation.

Bisbey, T. M., Kilcullen, M. P., Thomas, E. J., Ottosen, M. J., Tsao, K., & Salas, E. (2021). Safety culture: An integration of existing models and a framework for understanding its development. *Human Factors*, *63*(1), 88–110. https://doi.org/10.1177/0018720819868878

Bjertnaes, O., Iversen, H. H., Skyrud, K. D., & Danielsen, K. (2020). The value of Facebook in nation-wide hospital quality assessment: A national mixed-methods study in Norway. *BMJ Quality & Safety*, *29*(3), 217–224. https://doi.org/10.1136/bmjqs-2019-009456

Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. (2020). Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computer Methods and Programs in Biomedicine*, *188*, 105264. https://doi.org/10.1016/j.cmpb.2019.105264

Bleaney, G., Kuzyk, M., Man, J., Mayanloo, H., & Tizhoosh, H. R. (2018). Auto-detection of safety issues in baby products. In Mouhoub, M., Sadaoui, S., Ait Mohamed, O., & Ali, M. (Eds.). *Recent trends and future technology in applied intelligence* (Vol. 10868, pp. 505–516). Springer. http://arxiv.org/abs/1805.09772

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Boylan, A.-M., Turk, A., van Velthoven, M. H., & Powell, J. (2020). Online patient feedback as a measure of quality in primary care: A multimethod study using correlation and qualitative analysis. *BMJ Open*, *10*(2), e031820. https://doi.org/10.1136/bmjopen-2019-031820

Boylan, A.-M., Williams, V., & Powell, J. (2020). Online patient feedback: A scoping review and stakeholder consultation to guide health policy. *Journal of Health Services Research & Policy*, *25*(2), 122–129. https://doi.org/10.1177/1355819619870837

Carrillo-de-Albornoz, J., Vidal, J. R., & Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. *Plos One*, *13*(11), e0207996. https://doi.org/10.1371/journal.pone.0207996

Catino, M., & Patriotta, G. (2013). Learning from errors: Cognition, emotions and safety culture in the Italian air force. *Organization Studies*, *34*(4), 437–467. https://doi.org/10.1177/0170840612467156

Cornish, F. (2021). 'Grenfell changes everything?' Activism beyond hope and despair. *Critical Public Health*, *31*(3), 293–305. https://doi.org/10.1080/09581596.2020.1869184

Dixon-Woods, M., Suokas, A., Pitchforth, E., & Tarrant, C. (2009). An ethnographic study of classifying and accounting for risk at the sharp end of medical wards. *Social Science & Medicine*, *69*(3), 362–369. https://doi.org/10.1016/j.socscimed.2009.05.025

Doherty, C., & Stavropoulou, C. (2012). Patients' willingness and ability to participate actively in the reduction of clinical errors: A systematic literature review. *Social Science & Medicine*, *75*(2), 257–263. https://doi.org/10.1016/j.socscimed.2012.02.056

Dugan, J. R., & Breda, D. R. (1991). Complaints about police officers: A comparison among types and agencies. *Journal of Criminal Justice*, *19*(2), 165–171. https://doi.org/10.1016/0047–2352(91)90050-6

Entwistle, V. A., McCaughan, D., Watt, I. S., Birks, Y., Hall, J., Peat, M., Williams, B., Wright, J., & Involvement in Patient Safety Group. (2010). Speaking up about safety concerns: Multi-setting qualitative study of patients' views and experiences. *Quality and Safety in Health Care*, *19*(6), e33. https://doi.org/10.1136/qshc.2009.039743

Explosion. (2020). Release en_core_web_lg-2.3.1. GitHub. https://explosion/spacy-models/releases/tag/en_core_web_lg-2.3.1

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman Publishing Inc.

Frey, B., Buettiker, V., Hug, M. I., Waldvogel, K., Gessler, P., Ghelfi, D., Hodler, C., & Baenziger, O. (2002). Does critical incident reporting contribute to medication error prevention? *European Journal of Pediatrics*, *161*(11), 594–599. https://doi.org/10.1007/s00431-002–1055-0

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, *50*(1), 344–361. https://doi.org/10.3758/s13428-017–0875–9

Giardina, T. D., Haskell, H., Menon, S., Hallisy, J., Southwick, F. S., Sarkar, U., Royse, K. E., & Singh, H. (2018). Learning from patients' experiences related to diagnostic errors is essential for progress in patient safety. *Health Affairs*, *37*(11), 1821–1827. https://doi.org/10.1377/hlthaff.2018.0698

Gibbons, C., & Greaves, F. (2018). Lending a hand: Could machine learning help hospital staff make better use of patient feedback? *BMJ Quality & Safety*, *27*(2), 93–95. https://doi.org/10.1136/bmjqs-2017-007151

Gillespie, A. (2020). Disruption, self-presentation, and defensive tactics at the threshold of learning. *Review of General Psychology*, *24*(4), 382–396. https://doi.org/10.1177/1089268020914258

Gillespie, A., & Reader, T. W. (2018). Patient-centered insights: Using health care complaints to reveal hot spots and blind spots in quality and safety. *The Milbank Quarterly*, *96*(3), 530–567. https://doi.org/10.1111/1468-0009.12338

Goldberg, D. M., Khan, S., Zaman, N., Gruss, R. J., & Abrahams, A. S. (2020). Text mining approaches for postmarket food safety surveillance

using online media. *Risk Analysis*, Advance online publication. https://doi.org/10.1111/risa.13651

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Harnessing the cloud of patient experience: Using social media to detect poor quality healthcare. *BMJ Quality & Safety*, *22*(3), 251–255. https://doi.org/10.1136/bmjqs-2012-001527

Griffiths, A., & Leaver, M. P. (2017). Wisdom of patients: Predicting the quality of care using aggregated patient feedback. *BMJ Quality & Safety*, *27*, 110–118. https://doi.org/10.1136/bmjqs-2017-006847

Gu, G., Zhang, X., Zhu, X., Jian, Z., Chen, K., Wen, D., Gao, L., Zhang, S., Wang, F., Ma, H., & Lei, J. (2019). Development of a consumer health vocabulary by mining health forum texts based on word embedding: Semiautomatic approach. *JMIR Medical Informatics*, *7*(2), e12704. https://doi.org/10.2196/12704

Hald, E. J., Gillespie, A., & Reader, T. W. (2021). Causal and corrective organisational culture: A systematic review of case studies of institutional failure. *Journal of Business Ethics*, *174*, 457–483. https://doi.org/10.1007/s10551-020-04620-3

Herzer, K. R., Mirrer, M., Xie, Y., Steppan, J., Li, M., Jung, C., Cover, R., Doyle, P. A., & Mark, L. J. (2012). Patient safety reporting systems: Sustained quality improvement using a multidisciplinary team and "good catch" awards. *The Joint Commission Journal on Quality and Patient Safety*, *38*(8), 339–AP1. https://doi.org/10.1016/S1553-7250(12)38044-6

Howell, A.-M., Burns, E. M., Bouras, G., Donaldson, L. J., Athanasiou, T., & Darzi, A. (2015). 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*, *10*(12), e0144107. https://doi.org/10.1371/journal.pone.0144107

Howell, A.-M., Burns, E. M., Hull, L., Mayer, E., Sevdalis, N., & Darzi, A. (2017). International recommendations for national patient safety incident reporting systems: An expert Delphi consensus-building process. *BMJ Quality & Safety*, *26*(2), 150–163. https://doi.org/10.1136/bmjqs-2015-004456

Hulebak, K. L., & Schlosser, W. (2002). Hazard Analysis and Critical Control Point (HACCP) history and conceptual overview. *Risk Analysis*, *22*(3), 547–552. https://doi.org/10.1111/0272–4332.00038

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*, 216–225.

Jausan, M., Silva, J., & Sabatini, R. (2017). A holistic approach to evaluating the effect of safety barriers on the performance of safety reporting systems in aviation organisations. *Journal of Air Transport Management*, *63*, 95–107. https://doi.org/10.1016/j.jairtraman.2017.06.004

Kellogg, K. M., Hettinger, Z., Shah, M., Wears, R. L., Sellers, C. R., Squires, M., & Fairbanks, R. J. (2017). Our current approach to root cause analysis: Is it contributing to our failure to improve patient safety? *BMJ Quality & Safety*, *26*(5), 381–387.

King, A., Daniels, J., Lim, J., Cochrane, D. D., Taylor, A., & Ansermino, J. M. (2010). Time to listen: A review of methods to solicit patient reports of adverse events. *BMJ Quality & Safety*, *19*(2), 148–157. https://doi.org/10.1136/qshc.2008.030114

Kirkup, B. (2015). *The report of the Morecambe Bay investigation*. Morecambe Bay Investigation.

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Lane, J., Bhome, R., & Somani, B. (2021). National trends and cost of litigation in UK National Health Service (NHS): A specialty-specific analysis from the past decade. *Scottish Medical Journal*, *66*(4), 168–174. https://doi.org/10.1177/00369330211052627

Lawton, R., & Parker, D. (2002). Barriers to incident reporting in a healthcare system. *BMJ Quality & Safety*, *11*(1), 15–18. https://doi.org/10.1136/qhc.11.1.15

Leaver, M., & Reader, T. W. (2016). Human factors in financial trading: An analysis of trading incidents. *Human Factors*, *58*(6), 814–832. https://doi.org/10.1177/0018720816644872

Leveson, N., Dulac, N., Marais, K., & Carroll, J. (2009). Moving beyond normal accidents and high reliability organizations: A systems approach to safety in complex systems. *Organization Studies*, *30*(2–3), 227–249. https://doi.org/10.1177/0170840608101478

Levtzion-Korach, O., Frankel, A., Alcalai, H., Keohane, C., Orav, J., Graydon-Baker, E., Barnes, J., Gordon, K., Puopulo, A. L., & Tomov, E. I. (2010). Integrating incident data from five reporting systems to assess patient safety: Making sense of the elephant. *Joint Commission Journal on Quality and Patient Safety*, *36*(9), 402–410.

Lilford, R., Mohammed, M. A., Spiegelhalter, D., & Thomson, R. (2004). Use and misuse of process and outcome data in managing performance of acute medical care: Avoiding institutional stigma. *The Lancet*, *363*(9415), 1147–1154. https://doi.org/10.1016/S0140-6736(04)15901-1

Locock, L., Skea, Z., Alexander, G., Hiscox, C., Laidlaw, L., & Shepherd, J. (2020). Anonymity, veracity and power in online patient feedback: A quantitative and qualitative analysis of staff responses to patient comments on the 'Care Opinion' platform in Scotland. *Digital Health*, *6*, https://doi.org/10.1177/2F2055207619899520

MacLeod, G. (2018). The Grenfell Tower atrocity. *City*, *22*(4), 460–489. https://doi.org/10.1080/13604813.2018.1507099

Macrae, C. (2009). Making risks visible: Identifying and interpreting threats to airline flight safety. *Journal of Occupational and Organizational Psychology*, *82*(2), 273–293. https://doi.org/10.1348/096317908X314045

Macrae, C. (2016). The problem with incident reporting. *BMJ Quality & Safety*, *25*(2), 71–75. https://doi.org/10.1136/bmjqs-2015-004732

Makary, M. A., & Daniel, M. (2016). Medical error—The third leading cause of death in the US. *Bmj*, *353*, i2139. https://doi.org/10.1136/bmj.i2139

Marsh, C., Peacock, R., Sheard, L., Hughes, L., & Lawton, R. (2019). Patient experience feedback in UK hospitals: What types are available and what are their potential roles in quality improvement (QI)? *Health Expectations*, *22*(3), 317–326. https://doi.org/10.1111/hex.12885

Mazanderani, F., Kirkpatrick, S. F., Ziebland, S., Locock, L., & Powell, J. (2021). Caring for care: Online feedback in the context of public healthcare services. *Social Science & Medicine*, *285*, 114280. https://doi.org/10.1016/j.socscimed.2021.114280

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. *Journal of Management*, *44*(7), 2909–2933. https://doi.org/10.1177/0149206316657594

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *2*, 3111–3119.

Millman, E. A., Pronovost, P. J., Makary, M. A., & Wu, A. W. (2011). Patient-assisted incident reporting: Including the patient in patient safety. *Journal of Patient Safety*, *7*(2), 106–108. https://doi.org/10.1097/PTS.0b013e31821b3c5f

Mitchell, I., Schuster, A., Smith, K., Pronovost, P., & Wu, A. (2016). Patient safety incident reporting: A qualitative study of thoughts and perceptions of experts 15 years after 'To Err is Human.'. *BMJ Quality & Safety*, *25*(2), 92–99. https://doi.org/10.1136/bmjqs-2015-004405

National Academies of Sciences & Medicine. (2018). *Crossing the global quality chasm: Improving health care worldwide*. The National Academies Press.

NHS Improvement. (2018). NRLS official statistics publications: Guidance notes *(p. 16)*. NHS Improvement.

Ockenden, D. (2022). *Findings, conclusions and essential actions from the independent review of maternity services at the Shrewsbury and Telford Hospital NHS Trust*. House of Commons.

O'Connor, P., O'Dea, A., & Melton, J. (2007). A methodology for identifying human error in US Navy diving accidents. *Human Factors*, *49*(2), 214–226. https://doi.org/10.1518/001872007X312450

Öz, B., Özkan, T., & Lajunen, T. (2014). Trip-focused organizational safety climate: Investigating the relationships with errors, violations and positive driver behaviours in professional driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *26*, 361–369. https://doi.org/10.1016/j.trf.2014.03.004

Papanicolas, I., & Figueroa, J. F. (2019). Preventable harm: Getting the measure right. *BMJ*, *366*, l4611. https://doi.org/10.1136/bmj.l4611

Patel, S., Cain, R., Neailey, K., & Hooberman, L. (2015). General practitioners' concerns about online patient feedback: Findings from a descriptive

exploratory qualitative study in England. *Journal of Medical Internet Research*, *17*(12), e276. https://doi.org/10.2196/jmir.4989

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pfeiffer, Y., Manser, T., & Wehner, T. (2010). Conceptualising barriers to incident reporting: A psychological framework. *Quality and Safety in Health Care*, *19*(6), e60–e60. https://doi.org/10.1136/qshc.2008.030445

Placona, A. M., & Rathert, C. (2021). Are online patient reviews associated with health care outcomes? A systematic review of the literature. *Medical Care Research and Review*, *79*(1), 3–16. https://doi.org/10.1177/10775587211014534

Ramsey, L., Sheard, L., Lawton, R., & O'Hara, J. (2019). How do healthcare staff respond to patient experience feedback online? A typology of responses published on Care Opinion. *Patient Experience Journal*, *6*(2), 42–50. https://doi.org/10.35680/2372-0247.1363

Rashman, L., Withers, E., & Hartley, J. (2009). Organizational learning and knowledge in public service organizations: A systematic review of the literature. *International Journal of Management Reviews*, *11*(4), 463–494. https://doi.org/10.1111/j.1468-2370.2009.00257.x

Reader, T. W., & Gillespie, A. (2021). Stakeholders in safety: Patient reports on unsafe clinical behaviors distinguish hospital mortality rates. *Journal of Applied Psychology*, *106*(3), 439–451. https://doi.org/10.1037/apl0000507

Reason, J. (1998). Achieving a safe culture: Theory and practice. *Work & Stress*, *12*(3), 293–306. https://doi.org/10.1080/02678379808256868

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, *28*(1), 87–111. https://doi.org/10.1017/pan.2019.23

Sari, A. B.-A., Sheldon, T. A., Cracknell, A., & Turnbull, A. (2007). Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: Retrospective patient case note review. *BMJ*, *334*(7584), 79–82. https://doi.org/10.1136/bmj.39031.507153.AE

Shojania, K. G., & Thomas, E. J. (2013). Trends in adverse events over time: Why are we not improving? *BMJ Quality & Safety*, *22*(4), 273–277. https://doi.org/10.1136/bmjqs-2013-001935

Shwartz, M., Cohen, A. B., Restuccia, J. D., Ren, Z. J., Labonte, A., Theokary, C., Kang, R., & Horwitt, J. (2011). How well can we identify the high-performing hospital? *Medical Care Research and Review*, *68*(3), 290–310. https://doi.org/10.1177/1077558710386115

Stanton, N. A., Margaryan, A., & Littlejohn, A. (2017). Editorial: Learning from incidents. *Safety Science*, *99*, 1–4. https://doi.org/10.1016/j.ssci.2017.07.011

Stavropoulou, C., Doherty, C., & Tosey, P. (2015). How effective are incident-reporting systems for improving patient safety? A systematic literature review. *The Milbank Quarterly*, *93*(4), 826–866. https://doi.org/10.1111/1468-0009.12166

Sujan, M. A., Huang, H., & Braithwaite, J. (2017). Learning from incidents in health care: Critique from a Safety-II perspective. *Safety Science*, *99*, 115–121. https://doi.org/10.1016/j.ssci.2016.08.005

Toffolutti, V., & Stuckler, D. (2019). A culture of openness is associated with lower mortality rates among 137 english national health service acute trusts. *Health Affairs*, *38*(5), 844–850. https://doi.org/10.1377/hlthaff.2018.05303

Turk, A., Fleming, J., Powell, J., & Atherton, H. (2020). Exploring UK doctors' attitudes towards online patient feedback: Thematic analysis of survey data. *Digital Health*, *6*, https://doi.org/10.1177/2055207620908148

Turner, B. A. (1976). The organizational and interorganizational development of disasters. *Administrative Science Quarterly*, *21*(3), 378–397. https://doi.org/10.2307/2391850

Van Dael, J., Gillespie, A., Reader, T., Smalley, K., Papadimitriou, D., Glampson, B., Marshall, D., & Mayer, E. (2021). Getting the whole story: Integrating patient complaints and staff reports of unsafe care. *Journal of Health Services Research & Policy*, *27*(1), 41–49. https://doi.org/10.1177/13558196211029323

Vincent, C., Carthey, J., Macrae, C., & Amalberti, R. (2017). Safety analysis over time: Seven major changes to adverse event investigation. *Implementation Science*, *12*(1), 1–10. https://doi.org/10.1186/s13012-017–0695-4

Vincent, C., Neale, G., & Woloshynowych, M. (2001). Adverse events in British hospitals: Preliminary retrospective record review. *BMJ*, *322*(7285), 517–519. https://doi.org/10.1136/bmj.322.7285.517

Walker, G. (2017). Redefining the incidents to learn from: Safety science insights acquired on the journey from black boxes to Flight Data Monitoring. *Safety Science*, *99*, 14–22. https://doi.org/10.1016/j.ssci.2017.05.010

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, *87*, 12–20. https://doi.org/10.1016/j.jbi.2018.09.008

Waring, J. J. (2005). Beyond blame: Cultural barriers to medical incident reporting. *Social Science & Medicine*, *60*(9), 1927–1935. https://doi.org/10.1016/j.socscimed.2004.08.055

Weick, K. E., & Sutcliffe, K. M. (2011). *Managing the unexpected: Resilient performance in an age of uncertainty* (Vol. 8). John Wiley & Sons.

World Health Organization. (2017). *Patient safety: Making health care safer*. World Health Organization.

Zakkar, M. A., & Lizotte, D. J. (2021). Analyzing patient stories on social media using text analytics. *Journal of Healthcare Informatics Research*, *5*, 382–400. https://doi.org/10.1007/s41666-021-00097-5

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gillespie, A., & Reader, T. W. (2022). Online patient feedback as a safety valve: An automated language analysis of unnoticed and unresolved safety incidents. *Risk Analysis*, 1–15. https://doi.org/10.1111/risa.14002