How to get richness from health and safety data

Carlotta Vorbeck, Dominika Brzoska-Corenthy, James Thompson & Jodie Lewis

ERM, UK

SUMMARY

Gathering meaningful insights from data is a challenge faced by many organisations. In high-hazard industries data is crucial when it comes to identifying and understanding weak signals. These weak signals are important because they can indicate a problem and provide an opportunity for early intervention before an accident or incident occurs. As human factors specialists, when working with organisations to improve safety performance, our projects often involve review and interpretation of data. In this paper we share practical learnings and considerations at each stage throughout the data lifecycle, to maximise the insights that can be gained from health and safety data.

KEYWORDS

Data collection methods, weak signals, data analytics

Introduction

Research methodology is well established in academia but its application in high-hazard industries brings with it a set of unique challenges. Particularly in health and safety, where data can be used to save lives, research must be action oriented and data collected should be analysed and presented in as little time as possible.

As we discuss in this paper, setting up good data structure and collection methodology is essential in setting the health and safety industry up for success. We discuss the practical considerations of dealing with big data in multiple formats and share notes on how to present and structure the findings. We provide examples from work that we have completed for our clients to help visualise and explain the process of dealing with data in high-hazard industries.

This paper is organised in four data lifecycle steps, as summarised in Figure 1. In all these steps we focus on the value that the human factors discipline brings to understanding and getting richness from health and safety data.

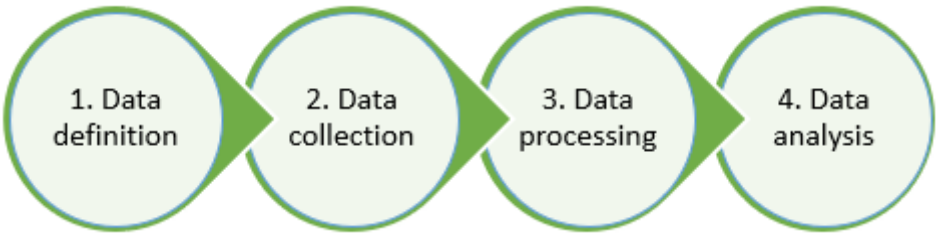

Figure 1: Key stages in the data lifecycle.

1. Data definition

When data is collected without a clear purpose, it is difficult to extract value from it. We have seen across all industries that our clients often collect and record huge amounts of data but struggle with putting this data into context or extracting real learnings from it. One may think that more data

equals greater insight, but this assumption is heavily dependent on a wide range of factors that ultimately tie back to understanding why the data is being collected at all.

Define the organisation's data needs

Before beginning to collect and process data, a critical first step is to define the organisation's needs and how data relates to them. From a safety perspective, these needs are defined by understanding the organisation's safety performance and implementing effective interventions to improve safety. As human factors specialists we also look to identify performance influencing factors (PIFs), so that we can understand the causes of human error and the impact on safety. Once the need is identified (why do we want to collect data?) it enables us to explore the hypotheses to be tested and the data that will be used to test them. Mapping out the data needs ensures that every data point is collected with purpose, as you cannot analyse and interpret data which was not yet collected. This process may seem obvious to the researchers among us, but it can be challenging to achieve in practice. In a high-hazard work environments there are a multitude of potential PIFs to consider and data analytics is only one component of a health and safety practitioner's diverse role.

We applied this approach in practice on a recent project with an organisation in the manufacturing sector. The company collected a vast amount of data but was not yet using this data to its full potential. Adopting a systems-thinking approach, we began by mapping out the potential PIFs, which included job/task factors, person factors and organisational factors. We then developed hypotheses to test the impact of these factors on safety outcomes. Existing data sources could then be linked to the hypotheses and any data gaps identified. Some examples of the hypotheses, PIFs, and data are shared in Table 1.

Table 1: Mapping hypotheses, PIFs and data variables.

Hypothesis	**Performance influencing factor**	**Data (Variables)**
Changes in production influence safety performance.	Production activities (job/task factor)	• Production volume/rates • Introduction of new product lines
Working patterns influence the likelihood of a safety incident.	Working patterns (organisational factor)	• Shift patterns • Working hours and overtime
Our training programs are delivering impact in terms of safety.	Competence and training (person factor)	• Training records / completion • Training needs assessments / matrix

2. Data collection

After determining the organisation's data needs, the next consideration should be the methodology for data collection. Applying a systematic approach to data collection will determine the level of depth, detail and amount of time and resource required for this process.

In any data-related project, it is helpful to consider the following:

1. Is an iterative or non-iterative data collection method needed?
2. Which data sources exist already? Is this data suitable and appropriately categorised?
3. Which data must still be collected, and which methodology would be best suited? Which skills are required for data collection?

By considering these three aspects of data collection you will be set up with a well thought out data collection methodology. We discuss each of these considerations in the following sections.

Iterative vs. non-iterative methodology

Where a project will stretch over a long period and can evolve over time, an iterative data-collection process will be most beneficial. Iterative processes are characterised by building, refining, and improving a project's approach over time until a satisfactory outcome is achieved. This allows for enough flexibility and targeted use of resources as priorities are developed or discovered and goals are achieved over time and in stages. The data collection plan should reflect this flexibility and provide multiple opportunities to re-assess priorities, change strategy, or formulate alternative hypotheses. Projects with time constraints, pre-defined end goals, and decisions to be made to progress will benefit most from a non-iterative process. Due to a lack of flexibility compared to iterative processes, outputs should be clearly defined and finalised at the beginning of the project.

Identify and review existing data sources

Health and safety projects can make use of existing data sets within the organisation, such as incident data, audits, employee hours, or length of service. However, there can be issues with the quality of pre-existing data that need to be carefully managed. Points to consider here are firstly the consistency of data collection, secondly the categorisation of data, and finally the validity of data and data sources. Where the data has not been collected in a consistent manner or format, the time necessary for data cleaning will be increased. The same is the case for data that has been poorly categorised or was not suitably categorised, as discussed in Section 3.

Collecting new data

Where new data must be collected for a project, much more time is required. However, this enables more flexibility and controlling of biases in the formulation of hypotheses or the identification of data sources. Typical data collection methods in health and safety related industries include surveys, interviews, site observations, and audits. It is important to ensure high validity when collecting data using these methods. When using surveys, you may collect both large text-responses or single-word responses. If responses contain terminology likely to affect later categorisation, it is best to consider using pre-categorised options to eliminate ambiguity around terminology. Pre-categorisation however comes with a risk of answers not matching any of the pre-selected answer options. This may be remedied by providing an option of 'other' to allow adding answers not considered by the researcher.

Whatever the method used, selecting the right individuals or groups to speak to when gathering data is pivotal. A diverse mix of job roles and demographics should be sought out, such as front-line workers, middle management, and safety professionals. Those who are directly affected by health and safety measures, for example frontline staff in the field, must be involved to gain comprehensive insights. Those indirectly affected, such as frontline supervisors, may also have valuable insights and perspectives of the workings of current social and procedural mechanisms. Failure to involve the right stakeholders due to personal bias or lack of consideration might mean that you only discover part of the picture, or that your data is incomplete. This will critically reduce any chances of drawing factual and meaningful conclusions from your data. The impact of psychosocial safety and the overall safety culture should also not be underestimated in determining the likelihood of uncovering critical and valid information and root causes for issues. The culture and implicit encouragement or discouragement around reporting and recording of incidents is heavily impacted by the perceived ability and safety of speaking up about health and safety issues in any industry.

Interviewing Skills

When speaking to individuals or groups during interviews, there are some guidelines we may borrow from interviewing methods in user research. When a problem is explored and relevant groups are familiar with this problem, targeted questions can be asked. This will help move conversations along to give interviewees the opportunity to share their perspective on the relevant issue. If the interviewer does not provide enough guidance in the conversation, what they are asking may be misunderstood and prevent the interviewees from sharing relevant information on the problem, causing you to lose valuable data. In cases where the root cause of a problem is not known yet, it is imperative to avoid asking leading questions as this will affect the interviewee's answers. They may believe that there is a specific answer you would like to hear or may stop talking once they feel they have answered your specific question. It is the interviewer's job to explore any possible leads without giving too much direction in the conversation to allow the interviewee to think out loud and make new connections within the information available to them. Monitoring one's own cognitive bias and effect on the conversation will enable you to find new information which was previously not considered at all. You may for example have a preconceived opinion that is causing you to dismiss certain causes of an issue because you find it obvious or illogical. You may also assume that an issue is due to a specific cause which you have previously come across in other conversations, something that can be described as 'tunnel vision'. However, in assuming that you have already found your answer, you may forego the real cause or other factors contributing to the problem which was simply not yet brought up. Human factors specialists are usually trained in these methods and are therefore well suited to carry out interviews in health and safety related industries.

Stakeholder Management

When collecting data, senior stakeholders should be included in the conversations early on to set expectations for the deliverables they will find most valuable for their department and the organisation. The research team may then manage expectations by explaining at a high level which analysis or methodology will be possible. For example, if the goal is to identify locations on site where most incidents are recorded, but senior stakeholders would like to make comparisons across sites in different countries, then the research team may advise that only sites with similar functions, geographies, layouts, etc. can meaningfully be compared to one another. This may determine how many separate comparisons must be made, which in turn may impact timelines, resource, and level of detail of the analysis. At this point, special care should be taken to be mindful of confirmation biases which may creep into the data collection philosophy. This may steer all data collection into a specific direction and cause important and valuable data points to be ignored, which in turn may lead to the core of the issue being missed completely. It is important to have a direction in mind to collect applicable data, but the data must provide a holistic view of a problem to identify root causes and weak signals of safety hazards, as this is what you should want to find.

Where senior stakeholders do not have a technical understanding of the data availability or the data format and tools required for the analysis later down the line, it is imperative that they include relevant technical specialists in conversations. Including them early on will enable effective and informed discussions about the project approach. Technical specialists may mean anyone who has a more thorough understanding of the data relevant to the project and underlying mechanisms at play. Recently, we encountered this problem during a project with a client in the mining sector. After deciding on comparing incident data of ten sites in a specific way, the client changed their approach to the analysis multiple times as they learned more details from their own technical specialists. Each time the approach was changed, the analysis team had to start their analysis from scratch due to the complexity of PIF categorisation at play. This increased the time spent on data analysis from an originally budgeted two days to over a week, which could have been avoided if technical specialists

had been involved at the beginning of the project to increase the senior stakeholders' understanding of the data and the subsequent best approach to comparing these ten sites to one another.

3. Data Processing

Before any data analysis can take place, the data must be checked for consistency. Where this was not considered or enforced during data collection, the data now needs to be cleaned and processed. The amount of time required for identifying, accessing, and understanding the content of different pre-existing data sets and their relationship to one another is also often underestimated. Setting aside enough time to understand these aspects will allow for a smooth start into data analysis.

Different names or abbreviations may have been used within the data to refer to a single location, task, function, or piece of equipment. These should be aligned to avoid the appearance of multiple factors having a small impact where in fact it is a single factor which has a large impact on health and safety. For numerical data, different units may have been used for the same metric between multiple data sets which must be corrected to allow for comparison. For example, one data source may record the rate of incidents for every thousand employee hours worked while another data source may record incident rates per million hours worked. Additional processing may also be required if numerical and text data are recorded within a single column of a database. For example, if a text description was entered alongside a numerical value within a single cell. These must be transformed into separate columns before analysis can take place.

Categorise the data

Categorising data is an important step in setting your data up for analysis. Grouping data allows for comparison of different factors within the same category or across categories. Analyses may then for example identify patterns of different factors interacting with one another, or simply identify the most influential factors in specific incident outcomes.

Quantitative data is often categorised by default, but qualitative data, such as free text, must often be categorised separately ahead of the analysis as part of data processing. Large text responses from surveys or incident descriptions must each be sorted into basic underlying categories or themes. These themes should capture the main PIFs discussed within the text. For example, the main theme of a text response could be summed up as 'work environment', 'ergonomics', or 'safety equipment'. These categories can then be used to perform numerical analyses of frequency counts for example. Where a single response discusses multiple themes, all themes should be considered. Appropriate explanations around multiple counts of PIFs or other responses should then be given in the report.

Data which is already sorted into categories, may require merging or further separation of categories to allow for meaningful comparison or contrasting. Many small, similar categories may be worth combining to give a better overall picture of the data. Conversely, a large but vague category can be broken down further to gain more detail in the analysis. For example, if the cause for a substantial amount of health and safety incidents is listed as 'human error', then richness can be added by breaking the category down into different types of human error.

Use the right tools to process the data

For initial processing of numerical or categorical data, developing charts and graphs is a way to visually check the data before beginning analysis. Additionally, pivot tables and filters can allow you to check and combine or contrast multiple factors efficiently and flexibly. For large open-text data fields, manual categorisation may not be practical, however the use of a simple Word Cloud may be sufficient to get across the key themes of the response. Quotes, which should always be anonymised, can be particularly powerful in highlighting specific findings compared to presenting numbers alone. When more detailed analysis is required, artificial intelligence (AI) machine

learning tools can be used to categorise text responses. In this case, a sample of these categorisations should be validated to help train the models before the outputs can be trusted. To avoid issues with data security, at ERM we developed our own in-house tools for H&S data categorisation. AI tools can also be used for translation of free text for multinational companies to enable creation of global datasets. Here again, when using AI, initial translation should be reviewed and adjusted to help train the AI models. Readily available alternatives, like Google Translate can be used but special attention should be paid to data security and translation of technical language, as conventional translation software is focused on casual conversation language, so technical terms and abbreviations are frequently mis-translated. Although training of a dedicated machine learning algorithm is a more costly method, it is in our opinion the most effective way to analyse health and safety data, considering the challenges of monitoring its continuous real-time development.

4. Data analysis

You may begin to see some patterns as the data is processed or even collected, but as mentioned previously it is best to wait until all data is processed to prevent 'tunnel vision' and limiting the analysis to these initial findings. If insights emerge which provide you with additional hypotheses, these can be noted as observations to be explored in the future, once data processing is complete.

Validate the hypotheses

Once the entire dataset is cleaned and relevant graphs are plotted, analysis can begin. It is good practice to focus initially on validating the original hypothesis established during data definition. It can be helpful to define expectations of the criteria which must be met by the data to either confirm or reject the hypothesis. This can reduce confirmation bias when analysing the data and interpreting the results, which can emerge from having a particular finding in mind that you want to find. These criteria should be set before looking at the data. If this was appropriately defined, it should be easy to identify data outputs which will support or reject the initial hypothesis. In any case, it is imperative to remember that correlation does not equal causation, meaning that even though two factors are impacting each other, this does not mean that one factor is causing the change in the other, but instead a third hidden factor may be causing changes in both of these factors because they are in some way related to each other. Certainly, for quantitative data appropriate statistical models have to be employed to draw accurate conclusions, and careful consideration of this principle and accurate root cause analysis is advised when dealing with qualitative data.

Accepting or rejecting your hypothesis is not likely to provide you with any breakthroughs or surprising insights about your organisation. It is, however, important to validate your initial assumptions about the organisation to build a strong foundation for your analysis. If the initial assumptions remain unverified and are in fact incorrect or present just part of a bigger picture, this might lead you to come to incorrect conclusions at the end of the analysis.

On occasion, the data collected might not support your initial hypothesis. In this case the analyst should ask themselves "why?":

1. What other conditions or events could cause the data to behave in this way?
2. Is the data collected not actually linked to my initial hypothesis?
3. Could there be two coexisting phenomena that cancel each other out?

These questions lead to creation of an alternative hypothesis. The alternative hypothesis may be validated by other supporting evidence. It is also possible that the data required to validate the alternative hypothesis was not collected. Although impractical, it is not always possible to identify all useful and necessary data points at the very beginning of the project. In these cases, new data points may be introduced to the dataset for future analysis, creating an iterative approach. In-field observations and follow up interviews might also help in validating the alternative hypothesis or

direct the analyst to new data points which might bring greater value to address the alternative hypothesis.

Identify trends and outliers

Once the initial and associated alternative hypothesis have been explored, trends over time may be identified from the data. Data may vary regularly over time, for example, spikes in the number of alarms at a certain time of the day or a specific day of the week. Data variation may also be continuous from a certain point in time, for example when a reportable issue was first noticed and then continuously reported thereafter. Correlations in data might also provide valuable insights, for example specific equipment frequently being named in reports related to specific events or injuries to specific body parts.

To improve your understanding of data variation and links, you should consider data which might not seem immediately related to health and safety, such as weather forecasts, specific national events, organisational or personal changes at important health and safety or technical positions, purchases of new equipment or important software updates, and so on. For example, could a spike in reporting be caused by a new policy? Is new software causing novel issues or is it alerting engineers to existing problems which were not picked up by the previous software and therefore never noticed? The analyst should also look at any anomalies in the data. Outliers should be identified and verified or corrected where appropriate, but never deleted without further investigation.

This is where the importance of presenting all data, including outliers, becomes apparent. What might initially look like an outlier might in fact be a weak signal of an event or condition happening at the site. When considered as evidence of a real issue, the outlier may reveal specific impacts on the organisation that are important to consider or correct. Depending on the time available, the outlier should be investigated further to explore the underlying conditions and if it may be a weak signal of a hidden health and safety hazard. Follow-up interviews and in-field observations might again provide valuable insights into the actual frequency and importance of the outliers.

Explore root causes

At any stage of the analysis, it is important to keep asking “why?”. Organisations can be quick to identify training as a corrective action, even when their training leading indicator is positive. It is frequently forgotten that training is not an effective way of preventing certain types of human errors such as lapses or slips. Instead, to identify appropriate controls to prevent such human errors, in-depth analysis should be carried out to uncover all associated PIFs. These PIFs could be personal, job/task related and organisational, and it is imperative to recognise that in most cases it is not one but a combination of different factors that lead to an outcome visible in the data. For example, vehicle collisions might be caused by driver distraction. This, however, does not necessarily mean that drivers are careless. Distraction could be caused by high number of alarms in the cabin. A small subset of the vehicle fleet may be suffering from a maintenance issue with the breaks, increasing the breaking distance. The analyst should, as much as possible, strive to uncover all causes of an issue present in data to identify any combination of factors which make up the real root cause. Only then can more effective controls than training, for example job/task-redesign, increased maintenance and equipment checks, or changes to shift patterns and staffing levels, be identified and implemented.

Manage stakeholder expectations

It is important to manage the stakeholders' expectations as well as your own. In less mature organisations, where thorough big data analysis has never been done before, the outcomes of the first analysis might point to various issues of which the organisation is already aware. These will

likely be frequent and severe enough to overshadow any weak signals. Strong signals should therefore be identified and addressed first before new, less obvious insights can be gained from the data. As the organisation matures, they can begin to look more closely at the data to uncover less obvious issues. In our experience this frequently presents a challenge as the funds for health and safety data analysis are often reduced when initial findings are reported and no groundbreaking, surprising information is included in the analysis. This lack of exciting brand-new insight may make it seem to organisations as though the initial investment into the analysis was not worth it, but it takes time and effort to discover and investigate weak signals.

Consider the audience when presenting findings

When presenting the findings, it is important to consider the audience and their individual goals and concerns. Different reports with different levels of detail should be presented to the senior management compared to middle-management or frontline staff. This nuanced approach will allow each audience to gain an understanding of the insights which are relevant to them, and thus develop actions more efficiently and appropriate to their own operations and targets. When presenting the results of your analysis, it is worth highlighting not only the safety critical hazards and root causes of problems that were identified, but to highlight the positive findings as well. The results presented to each audience should motivate them to implement future actions and solutions. Findings and graphs should be self-explanatory to de-couple them from the personal knowledge of the analyst. They should be easy to understand when viewed by stakeholders who were not initially involved in the project to allow wider sharing of the finding within the organisation. This enables a bigger impact of subsequent actions on the organisation as a whole.

Conclusion

At first glance the research steps presented in this paper may not seem to differ much from the traditional steps of research in academia. However, unlike in academia, only few people involved in a research project will have a strong foundation in data analysis, PIFs, or human behaviour. The latter frequently being the underlying cause of many modern-day industry accidents, incidents, and near misses. Human factors specialists have the unique ability to bridge this skill and knowledge gap as we are trained in data collection and analysis techniques, interviewing, and digging deeper when we see “human error” as the leading cause of industry problems. The aim of this paper was to highlight where these human factors skills can be the most helpful in data lifecycle.

Mindful of the value we, as human factors specialists, can bring to health and safety data analysis, we should also remember that other people involved in this process come from different backgrounds and bring their own unique skills and perspective to data gathering, processing and analysis. This is why it is important to set up the research expectations early on, engage with all the relevant stakeholders from the very beginning and involve them as much as possible at every step of the health and safety data lifecycle.

By outlining these considerations pertaining to the data lifecycle and research in health and safety, we hope to have shared some valuable learnings for organisations in safety critical industries. We hope that adopting these principles and recommendations will enable organisations to better understand the considerations required to obtain richness from their health and safety data. By adopting the practices we have outlined, they will be better equipped to detect weak signals, gain valuable insights into their health and safety performance, and enable impactful solutions to improve the safety of their operations and workforces.