Evidence Management for Compliance of Critical Systems with Safety Standards: A Survey on the State of Practice

Sunil Nair [a, *], Jose Luis de la Vara [a], Mehrdad Sabetzadeh [b], Davide Falessi [c]
[a] Certus Centre for Software V&V, Simula Research Laboratory, P.O. Box 134, 1325 Lysaker, Norway
[b] SnT Centre for Security, Reliability and Trust, University of Luxembourg, 4 rue Alphonse Weicker, L-2721, Luxembourg
[c] Fraunhofer Centre for Experimental Software Engineering, 5825 University Research Ct. Suite 1300 College Park, MD 20740, USA

Abstract

Context: Demonstrating compliance of critical systems with safety standards involves providing convincing evidence that the requirements of a standard are adequately met. For large systems, practitioners need to be able to effectively collect, structure, and assess substantial quantities of evidence.
Objective: This paper aims to provide insights into how practitioners deal with safety evidence management for critical computer-based systems. The information currently available about how this activity is performed in the industry is very limited.
Method: We conducted a survey to determine practitioners' perspectives and practices on safety evidence management. A total of 52 practitioners from 15 countries and 11 application domains responded to the survey. The respondents indicated the types of information used as safety evidence, how evidence is structured and assessed, how evidence evolution is addressed, and what challenges are faced in relation to provision of safety evidence.
Results: Our results indicate that (1) V&V artefacts, requirements specifications, and design specifications are the most frequently used safety evidence types, (2) evidence completeness checking and impact analysis are mostly performed manually at the moment, (3) text-based techniques are used more frequently than graphical notations for evidence structuring, (4) checklists and expert judgement are frequently used for evidence assessment, and (5) significant research effort has been spent on techniques that have seen little adoption in the industry. The main contributions of the survey are to provide an overall and up-to-date understanding of how the industry addresses safety evidence management, and to identify gaps in the state of the art.
Conclusion: We conclude that (1) V&V plays a major role in safety assurance, (2) the industry will clearly benefit from more tool support for collecting and manipulating safety evidence, and (3) future research on safety evidence management needs to place more emphasis on industrial applications.

Keywords: Safety-critical systems; safety certification; safety assurance; safety evidence; state of the practice.

---

*Corresponding author. Phone: +47 40 64 40 46, Fax: +47 67 82 82 01
E-mail: *sunil@simula.no* (Sunil Nair), *jdelavara@simula.no* (Jose Luis de la Vara), *mehrdad.sabetzadeh@uni.lu* (Mehrdad Sabetzadeh*), dfalessi@fc-md.umd.edu* (Davide Falessi)

Abbreviations:

ANSI	American National Standard Institute
AREMA	American Railway Engineering and Maintenance/of/way Association
ARP	Aerospace Recommendation Practice
BOM	Bill of Material
CAE	Claims, Arguments and Evidence
CENELEC	Comité Européen de Normalisation Electrotechnique (European Committee for Electrotechnical Standardization)
CS	Certification Specification
ECO	Engineering Change Orders
ECSS	European Cooperation on Space Standardization
FHA	Functional Hazard Analysis
FMEA	Failure Mode and Effect Analysis
FTA	Fault Tree Analysis
GSN	Goal Structuring Notation
IEC	International Electro-technical Commission
IEEE	The Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
JSP	Joint Service Publications
MIL STD	Military Standard
NAVAIR	Naval Air Systems Command
NORSOK	Norsk Sokkels Konkuranseposisjon
OHSAS	At Occupational Health & Safety Advisory Services
POEMS	Project-Oriented Environmental Management System
POSMS	Project-Oriented Safety Management System
RQ	Research Question
RTCA	Radio Technical Commission for Aeronautics
SLR	Systematic Literature Review
SPEM	System Process Engineering Metamodel Specification
SSA	System Safety Analysis
STANAG	Standardization Agreement
V&V	Verification and Validation

1 INTRODUCTION

Failures in safety-critical computer-based systems, including software-intensive ones, can have catastrophic consequences [1]. These systems are typically subject to *safety certification,* also referred to as *safety assurance*, as a way to ensure that the systems do not unduly harm people, property, or the environment. Safety certification is a stringent process, often conducted by an independent licensing or regulatory body, to provide an assurance that a system has met its stated safety properties, and that the system can be depended upon to deliver its intended service in a safe manner [2]. The safety criteria that need to be satisfied during certification are usually specified in the form of safety standards. Examples of safety standards include IEC61508 [3] for a wide range of electrical, electronic, and programmable electronic systems, DO-178C [4] for software in airborne systems, the CENELEC standards (e.g., EN 50129 [5]) for railway systems, and ISO26262 [6] for functional safety in the automotive domain.

Safety standards define requirements that a process or product needs to meet in order to be deemed safe. The system supplier has to demonstrate how these requirements are complied with by gathering convincing evidence during the system lifecycle. Safety evidence can be broadly defined as *"information or artefacts that contribute to developing confidence in the safe operation of a system"* [7]. Any artefact produced during a system's lifecycle may serve as evidence of a particular claim regarding system safety. In the context of certification and compliance with safety standards, safety evidence is also targeted at

showing the fulfilment of the requirements of a standard. Some generic examples of safety evidence, among several others, are testing results, system specifications, personnel competence, and source code.

For a realistically large system, practitioners need to collect and manage large quantities of safety evidence throughout the analysis, development, verification, maintenance, operation, and evolution of the system. This vast information has to be structured to show how it meets the requirements of a safety standard. If the evidence is not structured properly, its sheer volume and complexity can jeopardize the clarity of the satisfaction of the high-level safety objectives [8]. Safety evidence can be structured either graphically (e.g., with models) or textually.

As part of evidence management, practitioners must also assess the adequacy of the evidence. Adequacy is usually assessed based on the confidence in the information collected to support a particular claim about system safety [9]. Adequacy can be estimated qualitatively (e.g., via a confidence level) or quantitatively (e.g., via a numerical adequacy degree).

Traceability links are also usually required to capture the relationships between artefacts used as safety evidence. For example, a relationship exists between test cases and the requirements from which the test cases are derived. Due to the existence of these relationships, a change in one piece of evidence may affect others, possibly causing them to not be adequate anymore. For example, if a system requirement is modified, then the related test cases might have to be updated. The system supplier thus has to keep track of the various relationships in the body of evidence in order to be able to analyse change impact. This analysis aims at identifying the potential consequences of a change, or at estimating what needs to be modified to accomplish that change [10].

Although safety standards provide some guidance for managing safety evidence, they are generic and are typically large documents containing hundreds of pages and thousands of requirements [11]. For example, IEC 61508 – one of the most widely used safety standards – is organized into eight booklets (parts) with over 450 pages of text. For most safety standards, some degree of interpretation is required to tailor them to the context of application. This means that the system supplier has to decide based on the standard's guidance what type of evidence is best suited for a given scenario, and how it should be structured, assessed, and managed. Therefore, standards do not necessarily reflect industrial practices in safety evidence management, but only provide general information about practices that may be employed. This implies that the standards do not allow someone to know if certain practices are used, or to determine their frequency of use.

Despite the abundance of research focused on supporting and improving safety evidence management, few publications have been validated in real industrial projects or have provided empirical evidence about practices and perspectives in the industry. In a recent SLR on provision of safety evidence [7], we classified the publications selected based on the type of empirical validation that had been performed. The validation methods considered were case study, field study, action research, and survey. The SLR results showed that a vast majority of the publications (72%) had not been validated with any of these methods. Only a small fraction of the publications (17%) reflected on practices in actual projects, and even a smaller fraction (5%) had surveyed practitioners' activities and perspectives. In addition, the publications that had been empirically validated lack the degree of detail and rigor necessary to really understand the validation methodology and the level of generalizability to other contexts [12]. The number of data points of the publications was also very low, and most of the publications only related to a single application domain, standard, or organisation. As a result, very little knowledge exists about the global state of practice on safety evidence management.

The main objective of this paper is to contribute towards addressing the above gap by providing a general picture and new insights into practitioners' practices and perspectives regarding safety evidence management. Given the extensive research on the subject, it seems natural and of great importance to

analyse the perceptions of practitioners about the adoption and effectiveness of the existing tools and techniques for evidence management. For this purpose, an empirical study has been conducted in the form of a questionnaire-based survey [13]. The survey was targeted at practitioners who directly participated or had participated in evidence management for demonstrating the compliance of critical computer-based systems with safety standards. The content of the questionnaire was based primarily on the results of the above-cited SLR.

We obtained 52 valid responses from 11 different domains and 15 countries. We investigate the types of information and artefacts that are used as safety evidence and the techniques for structuring and assessing evidence. We further analyse practices for safety evidence change management and give insights into the current challenges that practitioners face in terms of safety evidence provision. In addition, we compare the results of the survey against the state of the art in order to identify major gaps and future research needs.

The survey represents a major step towards developing a better understanding of safety certification needs in practice, and its results can be useful both for academia and for industry. Researchers can identify gaps in the current state of the art that could be addressed in the future, as well as aspects in the state of the practice that might be improved by means of new research efforts. Practitioners can get a better understanding on how safety evidence can be managed according to the practices and perspectives reported. This can help them to adapt and ideally improve their own practices based on the way that other practitioners deal with safety evidence management. Furthermore, the evidence about the gap between research and practice was *anecdotal* until this study. We are not aware of any previous work that highlights this gap and its extent in an empirically rigorous manner. While further data collection would be beneficial for drawing stronger conclusions from our findings, the systematic procedure applied for conducting the survey combined with the high number and diversity of the respondents make us confident about the usefulness and representativeness of the results.

The rest of the paper is organized as follows. Section 2 presents the related work in the area. Section 3 describes the research method used in our study. Section 4 presents the survey results and our interpretation. Finally, Section 5 presents a summary of the results, our main conclusions, and future work.

2 Related Work

As mentioned above, we draw on the results of a SLR on the provision of evidence for safety compliance [7]. This SLR analyses 218 peer-reviewed papers published between 1990-2012, in order to (1) identify and classify the information and artefacts considered as evidence for safety certification, (2) determine the existing techniques for evidence structuring, (3) determine the existing techniques for evidence assessment, and (4) provide a list of challenges addressed for evidence provision. As a result of the review, a taxonomy of evidence types was provided, as well as categories of techniques for evidence structuring, of techniques for evidence assessment, and of challenges.

Out of the 218 publications selected, 61 had been validated by means of some empirical method and 37 presented insights into and thus evidence about industrial practices and perspectives. These publications correspond to action research (validation in real projects by the authors themselves; 26 publications), case studies (validation in real projects by practitioners different to the authors; 7 publications), or surveys (validation on the basis of practitioners' perspectives; 4 publications). One publication applied both action research and survey research [14]. Details of these publications can be found in [7].

When validating their work through surveys, a study reported the perspective on safety cases of ten practitioners from Swedish automotive companies [15]. Issues regarding audits of airborne software have been presented in [16]. Two studies surveyed the use of formal methods [14, 17], and one analysed

the experiences and opinions concerning tool qualification according to the RTCA DO-254 guidelines [2]. In another survey study, practitioners from Norway's oil and gas industry were asked about the use of the IEC61508 standard and their opinion about the application of model-based techniques to facilitate achieving compliance with this standard [18].

Related surveys can further be found in some European research projects. In the SafeCer project (http://www.safecer.eu), 19 partners completed a survey [19] and responded to questions about certification and development processes, component models, safety argumentation, and V&V practices. This project aims to provide support for system safety argumentation and for the generation of the corresponding evidence in a compositional manner for the automotive, avionics, construction equipment, and railway domains.

The study that we report in this paper has been performed in the context of OPENCOSS (http://www.opencoss-project.eu), an European research concerned with developing a common certification framework that spans the railway, avionics, and automotive domains in order to reduce certification time and costs via compositional and evolutionary certification. The OPENCOSS consortium consists of 17 partners from nine different European countries: three system manufacturers, one component suppliers, two quality assurance consultancies, five software tool vendors, one certification body, four research organizations, and one project management organisation. Within OPENCOSS, a baseline survey was previously conducted concerning the state of the practice in its consortium [20-23]. Responses were obtained from 15 partners on questions related to safety compliance management, safety case construction, cross-domain reuse of certification or assurance assets (such as evidence and evaluations), component reuse and modular certification, and practices involved in transparency of certification processes. With regards to the evidence management practices [23], partners indicated the general information included in certification document, how this information is structured and managed, and how traceability between documentation is managed.

While the above surveys provide a good starting point for understanding evidence management practices in the industry, the surveys focus mainly on the specific domains of the projects in which the surveys were conducted. These surveys do not provide a global picture of safety evidence management with adequate coverage of different domains. Furthermore, the results of the surveys were presented at a high level of abstraction, thereby lacking sufficient detail to understand concrete practices and viewpoints in the industry. For example, none of the existing surveys provide a detailed treatment of how practitioners assess the adequacy of evidence.

The survey in this paper fills these gaps by addressing a wider set of domains and providing more in-depth insights into the practice on safety evidence management in real-world settings. Furthermore, the study has the important advantage of building on the results of a recent state-of-the-art review. This has enabled us to conduct a systematic comparison between the state of the art and the state of the practice, which has not been possible in any of the above-cited surveys.

3 Research Method

We conducted a survey in order to provide insights into how practitioners deal with safety evidence management for critical computer-based systems. A survey is a comprehensive research method for collecting information to describe, compare, or explain knowledge and behaviour [13]. The investigation presented in this paper also corresponds to qualitative (also known as flexible) research. This type of research is mainly targeted at investigating and understanding phenomena within their real context and at seeking new insights, ideas, and possible hypotheses for future research [24].

Based on the guidelines for survey research presented in [13], the following subsections present the RQs, the survey design, instrument evaluation, data collection, data analysis, and threats to validity.

3.1Research Questions

The aim of the survey is to gain knowledge on how safety evidence is provided and managed by practitioners when having to demonstrate compliance with safety standards for critical computer-based systems. Within this scope, we formulated the following RQs.

- **RQ1. What types of information and artefacts are used as evidence for demonstrating compliance with safety standards?**
  The aim of this question is to determine the various information and artefacts provided, checked, or requested as evidence to demonstrate safety compliance and thus safety of a system.
- **RQ2. How is evidence change managed?**
  The aim of this question is to identify industrial practices for managing evidence evolution and performing evidence change impact analysis.
- **RQ3. What techniques are used for structuring evidence?**
  The aim of this question is to determine techniques that practitioners use for presenting evidence in order to show how it contributes to the fulfilment of the requirements of a safety standard.
- **RQ4. What techniques are used for assessing evidence?**
  The aim of this question is to identify types of techniques that are applied in industry for evaluating the confidence or adequacy of the evidence provided.
- **RQ5. What challenges do practitioners face for providing safety evidence?**
  The aim of this question is to identify problems that practitioners might face when having to provide safety evidence and thus to show compliance with safety standards.
- **RQ6. What gaps exist between the state of the art and the state of the practice regarding safety evidence management?**
  The aim of this question is to identify potential differences between the research reported in [7] and our findings about the practice. Consequently, we also intend to assess past research according to industrial practices and needs.

3.2Survey Design

We designed a cross sectional web-based survey [13], aimed at obtaining information from the participants at a fixed point in time based on their past experience in demonstrating compliance with safety standards. We created a structured questionnaire to collect data relevant to the RQs. The questionnaire can be found in [25].

Advantages and disadvantages of using online questionnaire-based survey have been studied in past research (e.g., [26]). We believe that a questionnaire-based survey is an effective way to address the RQs above, allowing us to: (1) measure many variables simultaneously; (2) reach a large number of experts all over the world with domain knowledge, expertise and experience in managing safety evidence for safety certification; (3) develop a representative picture of the attitudes and characteristics of a large population of experts who manage evidence for safety certification and assessment; (4) reduce potential bias from having, for example, only interviewed people involved in a single project.

The questionnaire was designed closely following the results of a large-scale SLR [7]. In its final version, the questionnaire had 21 questions and the expected time for completing it was around 15 minutes. While designing the questionnaire, we did not focus on any particular safety standard or domain and therefore we did not base the evidence requirements on, for instance, a single criticality level proposed by the standards. The aim of our study was instead to provide an overall and global picture of the state of the practice on evidence management without leaning towards any particular safety standard or domain.

The questionnaire began with a short introduction to the purpose of the study and details about the target population. The target population of the study corresponded to practitioners that directly participated or had participated in evidence management for demonstrating compliance of critical computer-based systems with safety standards. The practitioners can correspond to people who provide evidence (e.g., a component supplier), check evidence for others (e.g., a safety assessor), or request evidence (e.g., a certification authority).

In the next part, we collected background information about the participants related to the context in which they had participated in safety evidence management and their experience. Participants were then asked questions related to the RQs. Some parts were presented in randomized order. Further important highlights about the questionnaire are as follows:

- For the questions concerning the information and artefacts used as safety evidence, a list of 49 evidence types along with a short definition of each type was provided. The 49 evidence types correspond to the evidence taxonomy built as part of the SLR in our previous work [7]. In the taxonomy, the evidence is split into two main categories: (1) Process information, related to the process followed to develop and verify a system, and (2) Product information, related to a system itself (e.g., its design). Under Product information we further classified various testing evidence types.
- Respondents were asked to indicate the frequency of use for several evidence structuring and evidence assessment techniques with the help of a five-point frequency Likert scale adopted from [27]: *Never, Rarely, Sometimes, Very often, and Always*.
- Respondents were asked to rate the importance of 10 possible challenges for safety evidence provision using a five-point importance Likert scale adopted from [27]: *Unimportant*, *Of little Importance*, *Moderately Important*, *Important*, and *Very Important*.

Where possible, and since we did not ask about a specific project but rather the respondents' overall experience, the respondents were allowed to select more than one option in order to indicate that they had observed several practices. Respondents were also given the possibility to mention other options (e.g., other challenges), except for the questions in which we considered that no other options were really possible (e.g., Yes/No questions). The respondents were provided with a brief description of each question, a definition for each evidence type for common understanding, and examples for clarifying the possible answers to some questions. For instance, GSN was provided as an example of argumentation-based graphical notation for structuring evidence. Finally, an optional part for participation in follow-up studies was included at the end of the questionnaire.

3.3Instrument Evalution and Data Collection

A two-stage process was adopted to evaluate the survey instrument. First, the instrument was evaluated by a focus group in which three experts provided feedback. The three experts who evaluated the survey instrument are: (1) a safety assurance manager at a system manufacturer, (2) a product manager at a component supplier, and (3) a senior researcher on safety assurance. Each expert had at least 5 years of experience in safety-critical system development and safety certification. They evaluated the reliability and validity of the questionnaire, aiming at identifying any potential ambiguity in the questions posed. Some minor changes were made at this stage. In the second stage, a pilot study was performed with five practitioners (two safety assessors and three safety assurance managers). Each practitioner had more than 5 years of experience in safety-critical system development and safety certification. In addition to validating the understandability of the questionnaire, this process aimed to ensure that the time required for filling the questionnaire was within expectations. Based on the feedback received, some parts of the questionnaire were rephrased and some questions were removed.

The survey data was collected from August through November of 2012. The survey was distributed via two ways: (1) a social networking website, and (2) personal email invitations. First, the survey was advertised in a social networking website for people in professional occupations (http://www.linkedin.com). We joined several groups related to demonstration of compliance with safety standards and posted the survey in the discussions page. Some groups were related to system safety in specific application domains (aerospace, automotive, avionics, defence, medical, nuclear, oil and gas, and railway), whereas others were related to more general areas (e.g., embedded systems). We posted two remainders in one-month time. Secondly, after a month, we sent personal email invitations and subsequent reminders to some practitioners whom we knew and considered to be part of the target population of the survey. We further asked the recipients of our email invitations to publicise the survey to colleagues who could participate in the survey.

In total we obtained 80 responses. We rejected 28 of these due to being incomplete. Hence, a final set of 52 valid responses (65%) was obtained. By valid we mean that the respondents answered all questions and provided all the information to categorise them. Out of the 52 valid responses, 44 responses came during the first month after posting the survey on LinkedIn. We obtained other 8 valid responses after we sent out personal email invitations. These 8 responses could either be prompted by our invitations or because some LinkedIn group members submitted their response late, i.e., after a month. We do not know how many members of each group actually saw the survey advertisement. The number of members depends on factors such as how often they access the groups and whether they receive notifications about messages posted on the group pages. This information is not available to us.

Using social networking websites such as the one used in this paper as opposed to more traditional means (e.g., surveying a specific organization or direct invitation) has its advantages and drawbacks. These advantages and drawbacks are well-studied and have been elaborated in the empirical software engineering literature [28-30]. Some benefits of using social networks for data collection, especially when compared to direct and personal invitations, include: (1) increase in subjects' heterogeneity; (2) increase in the level of confidence in the representativeness of the sample; (3) increase in the number of potential respondents reached, and; (4) the possibility of reaching a population for which no centralized bodies of professionals exist. Our rationale behind advertising the survey on a social network as the main source of data was to try to (1) obtain a more global and heterogeneous sample so that the respondents represented different profiles (e.g., country, domain, standard, role, and experience) and (2) mitigate possible threats to validity arising from only collecting data from known or directly contacted practitioners (e.g., a less representative sample as a result of a lower ratio of responses from certain domains or countries).

3.4 Subject Characteristics and Data Analysis

We obtained valid responses from 11 different application domains with the highest number of respondents from the *Aerospace* industry, followed by the *Railway*, *Avionics*, *Automotive*, and *Defence*. Figure 1 shows the percentage and number (in brackets) of respondents that selected each application domain. When analysing the safety standards for which the respondents had provided, checked, or requested evidence for compliance, we identified a set of 32 different regulations or families of regulations (e.g., CENELEC standards for the railway domain). More than one safety standard was mentioned by 54% of the respondents. TABLE I presents the list of safety standards and regulations that were indicated in the study, their frequency (i.e., the percentage of respondents that mentioned them and their number in brackets), and a short description about the applicability of the standard. In relation to the country in which the respondents mainly work (Figure 2), we identified 15 different countries. Four respondents replied that they were involved in compliance with safety standards in multiple countries. As shown in Figure 3 (a), a large majority of the respondents were from *developer/manufacturer of final*

systems and *component/system supplier*. About 40% of the respondents had more than 10 years of experience in demonstrating compliance with safety standards (Figure 3 (b)), and about 71% of the respondents had participated in five or more projects (Figure 3 (c)).

When analysing data, we harmonized some responses based on the information provided by the respondents in the *"Others"* options of the questions. For example, one respondent mentioned animation when asked about product-based evidence. We regard this as *Simulation results* evidence, and thus modified the response accordingly.

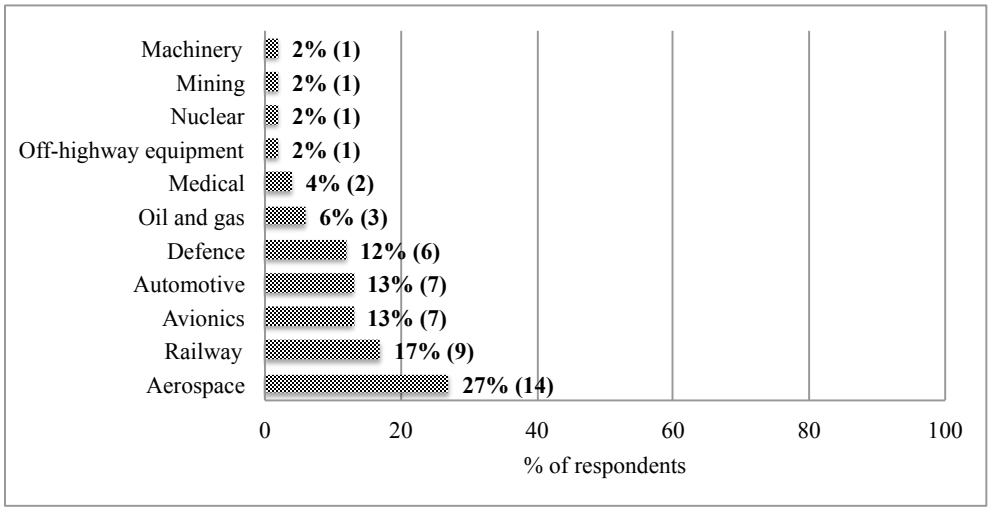


Figure 1. Application domains of respondents

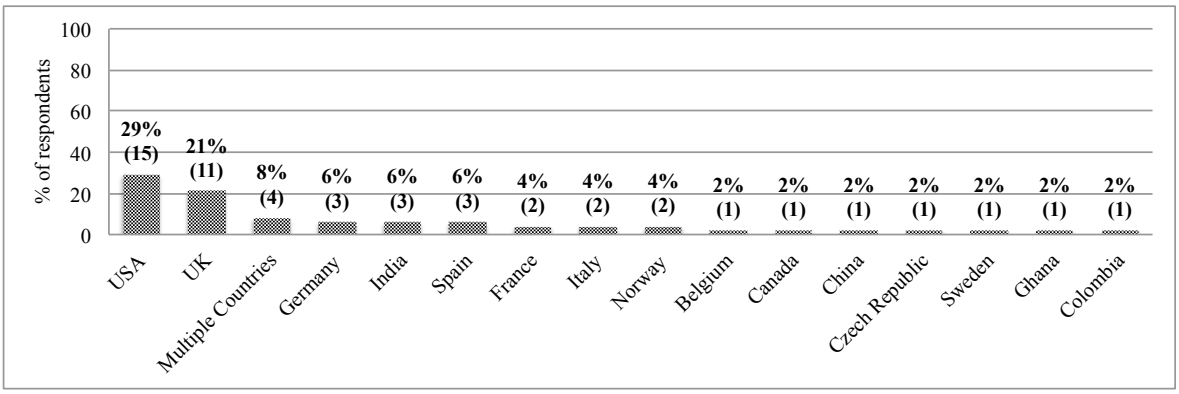
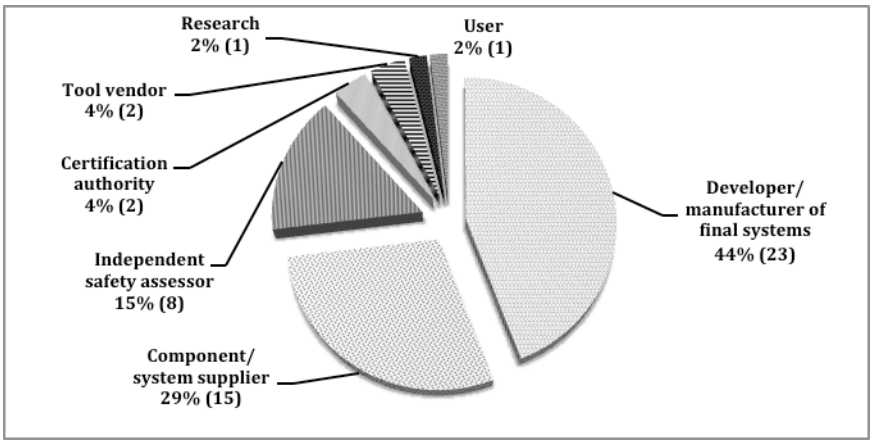


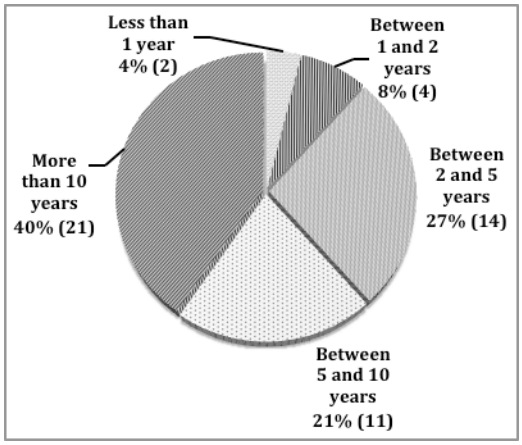
Figure 2. Countries of respondents

TABLE I. SAFETY STANDARDS MENTIONED IN THE RESPONSES

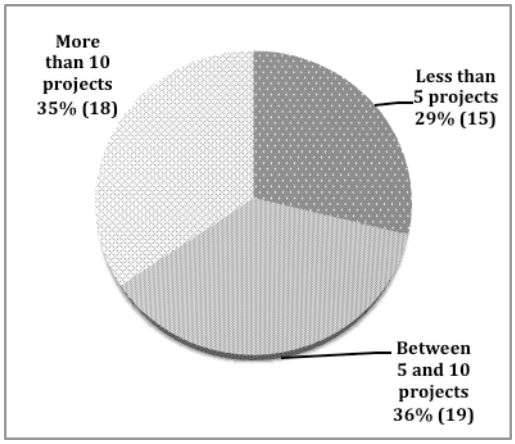
Safety Standard	Frequency	Description
RTCA DO-178B/C	33% (17)	Standard used for software consideration of commercial and military airborne systems and equipment
CENELEC Standards	19% (10)	Set of standards (EN50126, EN50128, and EN50129) for railway safety across Europe
IEC 61508	15% (8)	Standard used for the certification of electrical, electronic, or programmable electronic systems
ISO 26262	13% (7)	Standard for functional safety of road vehicles
MIL-STD-882	12% (6)	Standard for system safety in US military
UK Def Standards 00-55/56	10% (5)	Standard established by the Ministry of Defence (MOD) in the UK for providing safety management requirements for defence systems
RTCA DO-254	8% (4)	Standard that provides guidance for the development of airborne electronic hardware
ARP 4754	6% (3)	Aerospace recommendation practice for the development and certification of aircraft systems
IEC 62304	4% (2)	Standard that specifies lifecycle requirements for the development of medical software and software within medical devices
IEC 60601	4% (2)	Series of technical standards for the safety and effectiveness of medical electrical equipment.
ARP 4761	2% (1)	Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment
ISO 14971	2% (1)	Standard that establishes the requirements for risk management to determine the safety of a medical device
OHSAS 18001	2% (1)	A British standard for occupational health and safety management systems to help all kinds of organizations put in place demonstrably sound occupational health and safety performance
AREMA	2% (1)	The American Railway Engineering and Maintenance-of-way Association publishes standards and offers guidelines and best practices for railway engineering
IEC 61513	2% (1)	Application of IEC61508 to the nuclear industry
ISO 10993	2% (1)	A series of standards for evaluating the biocompatibility of a medical device prior to a clinical study
NORSOK	2% (1)	A set of standards aimed to ensure adequate safety, value adding, and cost effectiveness for petroleum industry developments and operations.
ANSI/ISA-84.00.01-2004	2% (1)	Standard that provides guidance on the specification, design, installation, operation and maintenance of safety instrumented functions
ISO 15998	2% (1)	Standard that specifies performance criteria and tests for functional safety of machine-control systems using electronic components in earth-moving machinery and its equipment
JSP 454	2% (1)	MOD Joint Service Publications that define the policy and identify specific regulatory requirements for system safety and environmental assurance for land systems.
POEMS	2% (1)	Project-oriented environmental management system manual that identifies the significant potential environmental impacts and risks associated with equipment systems and services acquisition projects
POSMS	2% (1)	Project-oriented safety management system that describes the safety management processes and procedures to be employed during a project's life cycle by defence equipment and support, and contractors working for them
Military Aviation Authority Regulation	2% (1)	Part of the MOD regulations, it is responsible for the regulation, surveillance, inspection, and assurance of the defence air operating and technical domains
ISO 13849	2% (1)	Standard that provides safety requirements and guidance on the principles for the design and integration of safety-related parts of control systems, including the design of software
RTCA DO-160	2% (1)	Standard for environmental test of avionics hardware
ECSS-E-ST-40C, ECSS-E-ST-80C	2% (1)	Series of software-related standards intended to be applied together for the management, engineering, and product assurance in space projects and applications
STANAG 4671	2% (1)	Standardization agreement from the NATO Standardization Agency that contains a set of technical airworthiness requirements intended primarily for the certification of fixed-wing military unmanned aerial vehicle systems
NAVAIR 13034	2% (1)	Standard that establishes policy, responsibilities, and procedures for executing airworthiness reviews resulting in Naval Air Systems Command flight clearances for all Department of Navy air vehicles and aircraft systems.
AMC 1303	2% (1)	It is a set of certification specifications for very light airplanes
CS-25.1309	2% (1)	Certification specification for large airplanes
IEEE 12207	2% (1)	Standard that establishes a common framework for software life cycle process.
Joint Software System Safety Engineers Handbook	2% (1)	Handbook that provides management and engineering guidelines to achieve a reasonable level of assurance that a piece of software will execute within the system context with an acceptable level of safety risk



(a)



(b)



(c)

Figure 3. Respondents' (a) organization role, (b) years of experience and (c) number of projects

3.5Threats to Validity

In this section, we discuss the validity threats to our study and how they were mitigated. The four perspectives presented in [31] are used as a reference.

Construct validity: This type of validity is concerned with the relationship between a theory behind an investigation and its observation. We guaranteed confidentiality and anonymity of the responses and allowed the respondents to complete the survey without identifying themselves in order to mitigate potential problems of evaluation apprehension. The threat of providing an incomplete list was mitigated by giving an option to mention additional information (*"others"* option) when considered possible. In each questionnaire part, respondents were reminded to answer the questions in relation to the application domain selected. Obtaining data from a set of respondents with different backgrounds mitigated mono-operation bias.

Conclusion validity: This type of validity is concerned with the relationship between a treatment and its outcome. To make the respondent familiar with the context of the study and its purpose, we provided an introduction to the survey and introductions to its different parts. To mitigate threats of misunderstanding the survey questions, we provided the respondents with information about the intent of the questions and definition of the terminology used. The definitions were based on existing definitions in the literature and from the results of the SLR. Instrument evaluation allowed us to mitigate

ambiguity and misinterpretation and to validate the survey description. The order of presentation for the different parts, questions, and options of the questionnaire were randomized where possible. This mitigated the threats to omission of questions due to fatigue.

Internal validity: This type of validity is concerned with the causal relationship between a treatment and its results. Developing the survey instrument with close relation to a SLR mitigated threats of instrumentation. Moreover, several experts had validated the taxonomy of evidence discussed in the SLR, which makes us believe that it represents the closest available perspective of the practitioner's understanding and needs. In addition, none of the respondents mentioned any new evidence type that was not represented already in the taxonomy. The use of well-established Likert scales minimized threats related to the elicitation of expert opinions. Performing the pilot study and a focus group discussion also helped in mitigating instrumentation threats. Designing the survey instrument so that it could be completed in approximately 15 minutes helped mitigate maturation and mortality. Randomizing most of the parts of the survey also mitigated maturation in specific questions and options. Despite the fact that 27 people (those who did not answer all the required questions) can be considered to have dropped out, we think that mortality did not affect the study based on the heterogeneous background of the valid responses.

External validity: This validity is concerned with the generalization of the conclusions of an investigation. The study was aimed at characterizing and understanding the state of practice in safety evidence management in industry. It also corresponds to qualitative research and is not meant to generalize its conclusion beyond its context. However, understanding the phenomena under study might help in understanding other cases. The survey was advertised in a social networking website to different groups interested in different application domains. This contributes to external validity by enabling us to collect responses from a diverse pool of respondents. In this sense, no domain, standard, or country was selected by more than 33% of the respondents, indicating the absence of heavy bias towards a particular domain, standard, or country. There are also two other aspects that make us confident about the validity and representativeness of our sample. First, the subject characteristics are in line with the results of our previous SLR. For example, (1) avionics, aerospace, automotive, and railway were the four domains most frequently found, (2) UK and US were the two countries whose institutions had published a higher number of papers, and (3) DO-178 was the standard most frequently found. Second, The subject characteristics are also in line the characteristics of LinkedIn groups. For example, the domain-specific group in which the survey was advertised with the higher number of members was on aerospace, and the standard-specific group was on DO-178. We consider that there exist a correlation between the number of members of LinkedIn groups on a specific area and the number of practitioners in that area. LinkedIn users must request to join a group, and the groups over which the survey was advertised are primarily concerned with practice-related aspects (e.g., how to address a requirement from some specific safety standard). We therefore assume that most of the members were practitioners.

4 Results And Discussion

This section presents the results of the survey and how we interpret them. A subsection has been created for each RQ.

4.1RQ1: What Types Of Information And Artefacts Are Used As Evidence For Demonstrating Compliance With Safety Standards?

Figure 4 shows the 16 process-based evidence types provided as options in the questionnaire in the vertical axis, and the percentage and number (in brackets) of respondents who selected each type in the horizontal axis. *V&V plan* was the most recognized process-based evidence type. The second most selected type was *Development plan*, followed by *Safety management plan* and *Configuration*

management plan. Only four process-based evidence types were selected by less than 50% of the respondents: *Operator competence specification*, *Communication plan*, *Reused component historical service data*, and *Development and V&V staff competence specification*.

As for the product information category, shown in Figure 5, we identified that *Requirements specification* was the most selected product-based evidence type. The second most selected type was *Test results*, followed by *Test case specification* and *Design specification*. The least identified evidence type in the product information was *Theorem proving results*. Other product evidence types selected in low percentages were *Model checking*, *Object code*, *System historical service data*, and *Accidents specification*. These four types were selected by less than 50% of the respondents.

Since the *Testing results* evidence type is a very broad category, we decomposed it into 16 finer-grained types, shown in Figure 6. As indicated in Figure 6, we identified that *System testing* was the most selected type in this category, followed by *Functional testing*, *Normal range testing*, and *Acceptance testing*. The least selected testing type was *Non-operational testing*. All the other testing types were selected by more than 50% of the respondents.

We did not find any new evidence types mentioned in the *others* sections by the participants.

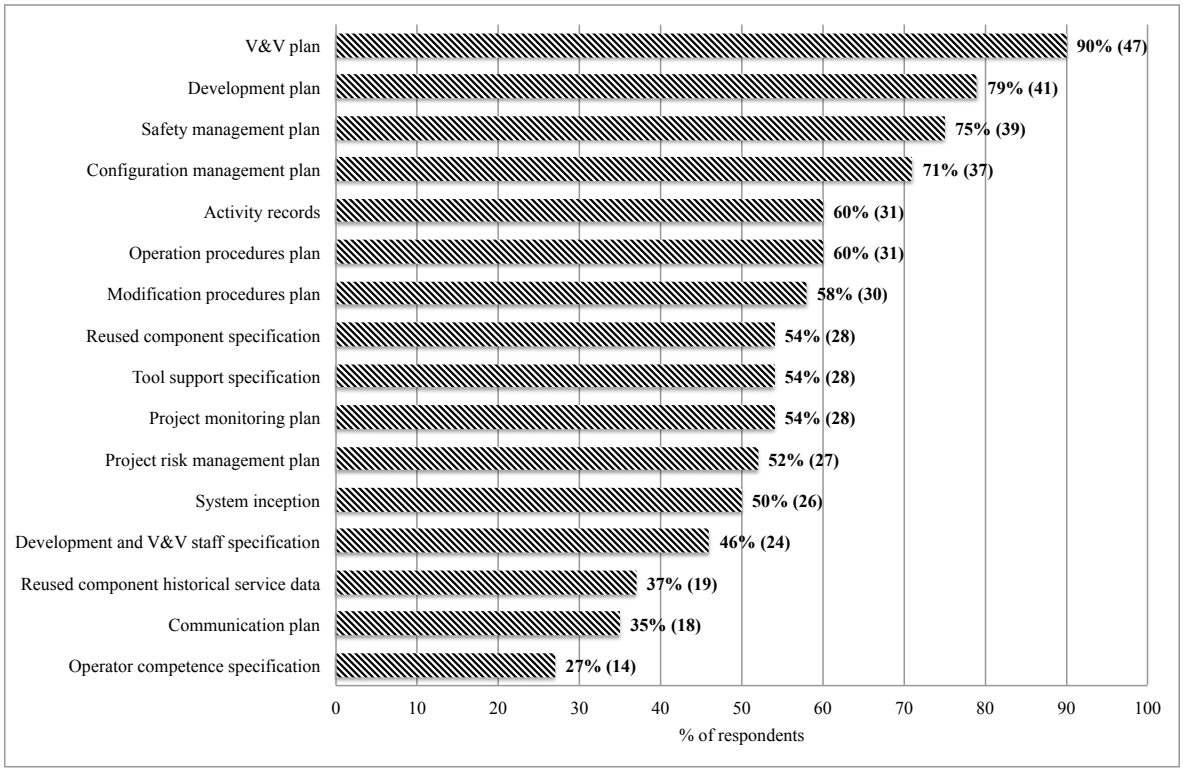


Figure 4. Frequency of process evidence types

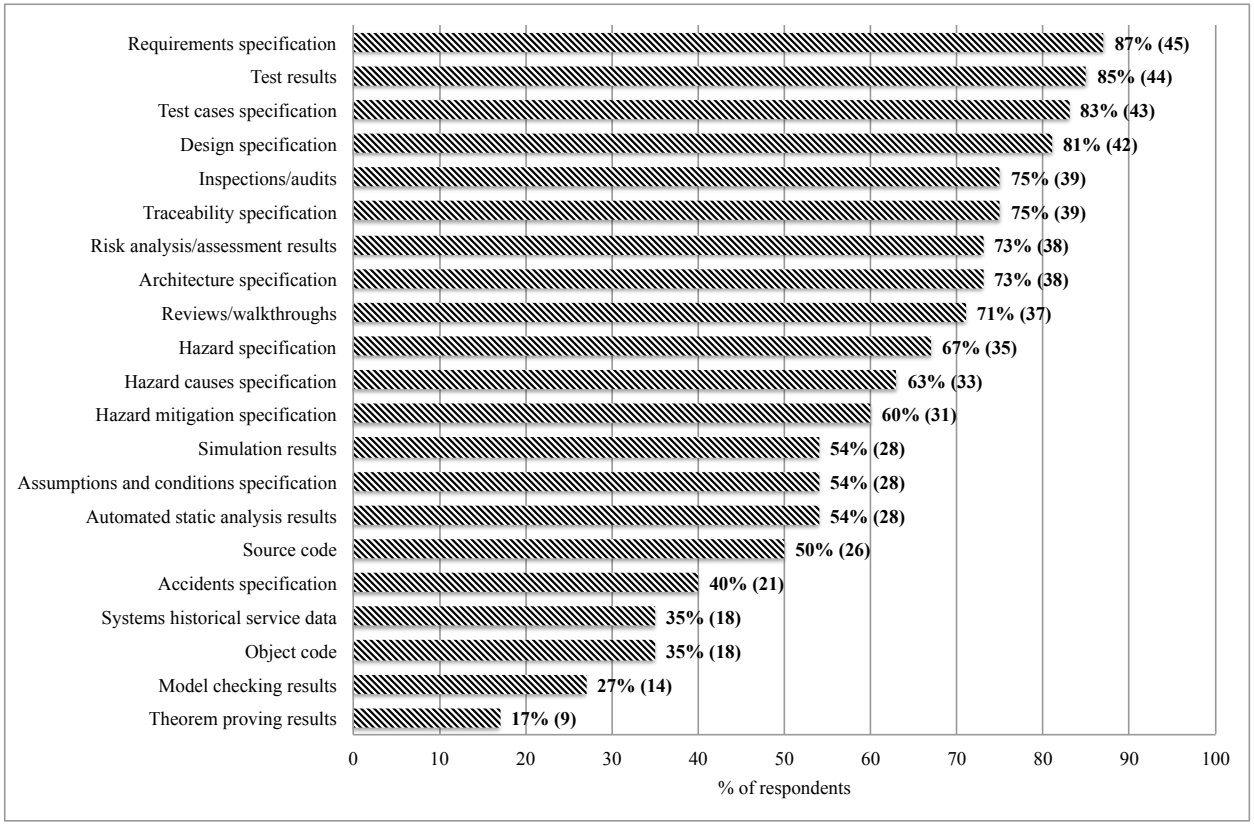


Figure 5. Frequency of product evidence types

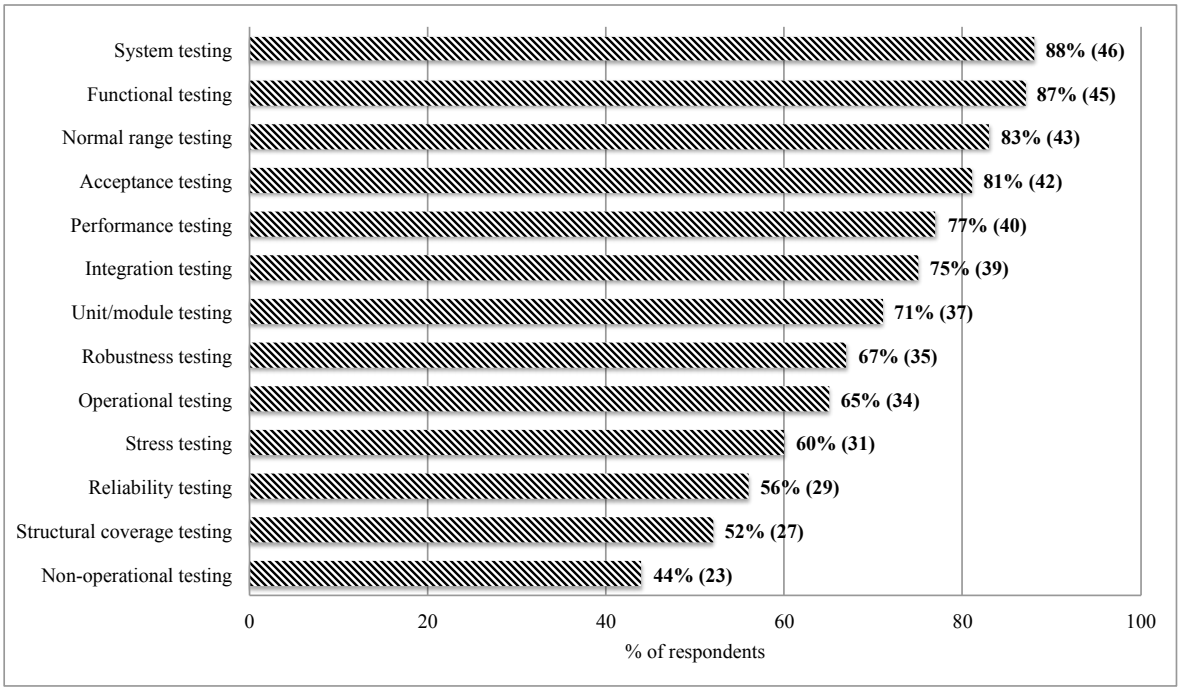


Figure 6. Frequency of testing evidence types

We have identified in this study that V&V-related evidence types such as *Test results, Test case Specification,* and *V&V plan* have been very often reported evidence types. Results from previous studies [23, 27] also show that high importance is given to the testing and verification process of a

safety-critical system for its certification. Consequently, and in general, these types seem to be among the ones with a greatest relevance for compliance with safety standards. Nonetheless, *Requirements specification, Design specification,* and *Development plan* (selected by more than 40 respondents) also seem to have a major role.

Based on the results, we think that there are several aspects that might require further analysis in future research. For example, future studies could analyse (1) when and why an evidence type with a purpose similar to another is selected (e.g., *Inspections/audits* instead of *Reviews/walkthroughs*), or when and why they are combined, and (2) if the lower selection of *Reuse component historical service data* in relation to *Reused component specification* implies that past operation is not a major aspect when having to show component safety (e.g., this might apply to real-time operating systems). We are also intrigued by the fact that evidence types concerning hazards and risks are not among the most frequently reported product-based types. A plausible and likely answer could be that such information is embedded in *Requirements specification* (e.g., in the form of safety requirements or measures).

4.2RQ2: How Is Evidence Change Managed In Practice?

The percentage and number (in brackets) of responses for ways to check the degree of evidence completeness is shown in Figure 7. Most of the respondents indicated that the degree of completeness for the evidence is checked manually (e.g., using paper-based checklists). A majority of the respondents (79%; 41 respondents) also noted that they provide, check or request details about how the change of a piece of evidence has affected other pieces of evidence. When asked about how they analyse the effect of the change of a piece of evidence on other pieces, 46% of the respondents noted manual checks according to some predefined process. Approximately the same percentage of respondents replied that the effect is checked manually although without following any predefined process. One respondent mentioned the use of modular software safety cases [32]. Figure 8 shows the frequency of the evidence change effect techniques.

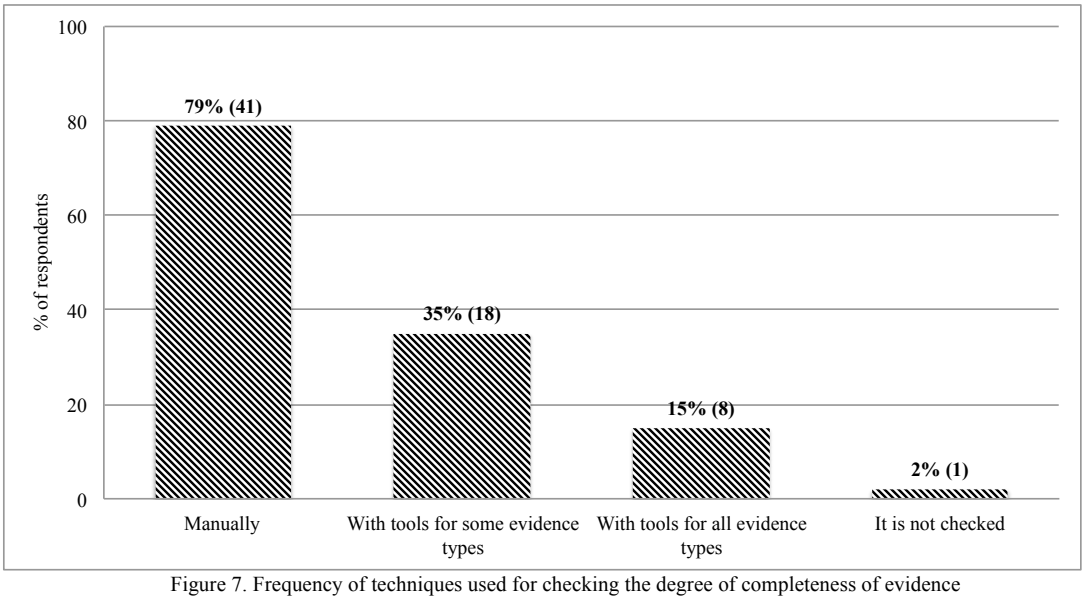


Figure 7. Frequency of techniques used for checking the degree of completeness of evidence

The majority of the respondents indicated that *Traceability matrices* are used for capturing the traceability between different pieces of evidence that they provide, check or request, whereas almost a fourth of them indicated the use of *Models*, *Hyperlinks,* or some *Naming conventions*. Frequency of

response to this question is shown in Figure 9. Some respondents provided additional information about practices for recording traceability. Single respondents acknowledged the use of ECOs [33], BOMs [34], Excel Spreadsheets, text documents created by version control tools and standard document templates, and safety analysis techniques like FTA, FMEA, FHA and SSA [35]. Three respondents mentioned IBM DOORS to record traceability information. Another respondent indicated that traceability information is normally embedded in a variety of documents, which combines one or more of the techniques proposed in the list (*Models*, *Matrices*, etc.), and that usually constraints on effort and cost lead to less comprehensive traceability.

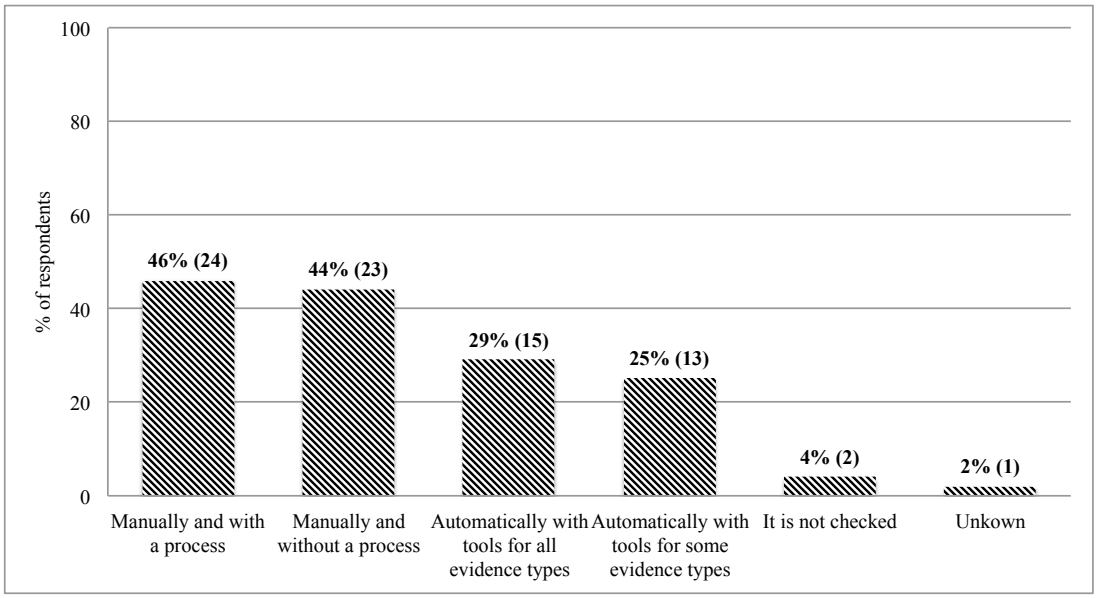


Figure 8. Frequency of techniques used for checking the effect of evidence change

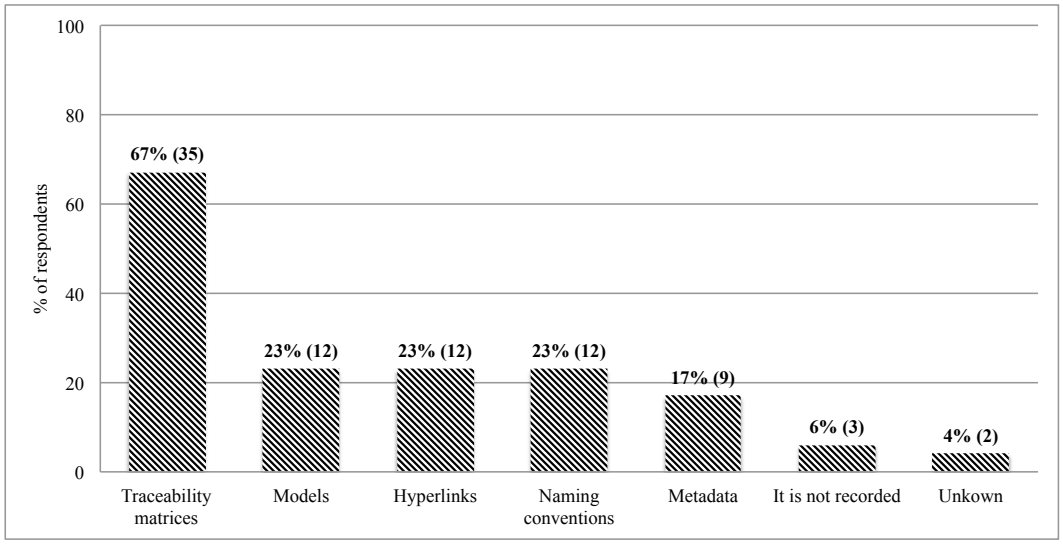


Figure 9. Frequency of evidence traceability recording techniques

When comparing the results obtained with previous surveys, we identify that the results are inline in general. For example, the results in [23], which was performed in a more limited setting with 15 partners from the OPENCOSS project, also suggest the use of traceability matrices as the most common technique for recording traceability in safety certification documents. DOORS was reported in [19] as a

tool for recording traceability of safety-critical systems. Nonetheless, it must be noted that most of the phenomena analysed in our survey (e.g., techniques used for checking evidence completeness or change effect) had barely been studied before or not studied at all.

An especially relevant finding is that the results suggest that evidence change management is mainly performed manually. Given the complexity of such an activity and the importance of executing it adequately, it seems that industry would benefit from more tool support. It could also be further analysed why practitioners do not use more tool support for this activity. Some possible reasons could be the lack of really suitable tools or the existence of factors that hinder their adoption (e.g., costs or training required). Another interesting finding is the fact that only 25% of the respondents did not select *Traceability specification* as a product-based evidence type, whereas only one respondent indicated that traceability is not recorded. In our opinion, this means that practitioners are concerned about the need for keeping traces regardless of whether they have to provide them as safety evidence. Consequently, there must exist stronger reasons for traceability other than compliance for some practitioners. One such possible motivation might be to perform change impact analysis in order to identify the impacted areas and take mitigation steps. It might also be studied in the future why practitioners might not need to check evidence completeness, analyse change impact, or record traceability. Nonetheless, a reason for obtaining these results in the survey might simply be, for instance, that the respondents (and thus the projects in which they had participated) had a limited scope, or were concerned only with some specific activity such as programming. Therefore, these aspects simply did not apply to them.

4.3RQ3: What Techniques Are Used For Structuring Evidence In Practice?

TABLE II shows the frequency of use of different evidence structuring techniques, indicating the total number of responses (N) for each technique, their median, and their mode (in bold). Except *Process models* such as SPEM and *Argumentation-based graphical notation* such as the GSN, the median of the techniques as used in practice is *Sometimes*. *Process models* and *Argumentation-based graphical notations* are the only techniques whose mode is *Never*, whereas *Textual templates* and *Structured text* have the highest modes (*Very Often*). *Textual templates* is also the technique most frequently reported as being used *Always*, as well as the technique reported as used by the highest number of respondents (91.7%). Therefore, the results suggest a generalised and frequent use of *Textual templates* for structuring evidence.

TABLE II. FREQUENCY OF USE OF EACH EVIDENCE STRUCTURING TECHNIQUE

Evidence Structuring Technique	N	Median	Never	Rarely	Sometimes	Very Often	Always
Textual templates	*49*	*Sometimes*	*8.3% (4)*	*22.4% (11)*	*18.4% (9)*	***34.7% (17)***	*16.3% (8)*
Structured Text	*49*	*Sometimes*	*20.4% (10)*	*8.2% (4)*	*26.5% (13)*	***38.8% (19)***	*6.1% (3)*
Conceptual/information models	*50*	*Sometimes*	*18% (9)*	*16% (8)*	***36% (18)***	*22% (11)*	*8% (4)*
Unstructured text	*49*	*Sometimes*	*14.3% (7)*	*22.4% (11)*	***32.7% (16)***	*26.5% (13)*	*4.1% (2)*
Argumentation-based graphical notations	*49*	*Rarely*	***36.7% (18)***	*14.3% (7)*	*20.4% (10)*	*24.5% (12)*	*4.1% (2)*
Process models	*46*	*Rarely*	***32.6% (15)***	*30.4% (14)*	*17.4% (8)*	*13% (6)*	*6.5% (3)*

Some respondents mentioned additional techniques to structure evidence: FTA and FMEA (one respondent), and DOORs (two respondents). This is in line with the responses to how traceability is recorded. One respondent mentioned the use of a wide set of systems for DO-178B and DO-254 compliance, consisting of Compliance Management System, Document Review Management System, Electronic File Management System, Reviews and Analysis Management System, Requirements Management System, Problem Reporting Management System and Workflow, and Coverage Analysis Management System. This response shows the complexity that evidence structuring can entail in

practice for complex systems, as practitioners can have to deal with a wide range of evidence types and supporting tools.

Previous work has also acknowledged the use of textual templates documentations for structuring evidence [18], although it did not indicate its overall frequency. Another survey [20] reports the use of *Argumentation-based graphical notations* such as GSN and CAE for structuring claims, arguments, and evidence as most popular, but our results note differences in the practice. Basically, the fact that these graphical notations are the most popular ones for argumentation does mean that *Argumentation-based graphical notations* are widely used in practice. Although promising results in the use of models for structuring and managing evidence have been reported in [18], it seems that such approaches are not extensively used in industry yet. Nonetheless, this makes sense to use because the use of models for evidence structuring has been proposed recently. Industry might also have been using some evidence structuring techniques for decades, without considering to adopt other techniques or being aware of them. The scope of the related work (in terms of the countries from which the respondents are) might be a possible explanation for the differences with the results of our survey too.

An aspect that could be the source for new research efforts is how practitioners show process compliance, and probably more interestingly how third parties request its demonstration. The results suggest a low use of process models despite the fact that they are targeted at, for instance, facilitating communication. It would be interesting to study if the use of models and graphical notations really provides benefits for demonstration or management of compliance with safety standards, and if these benefits could not be obtained by means of text-based approaches. Another open question is the purpose of using the model-based techniques in TABLE II, as the ratio of respondents indicating the use of models for traceability (Figure 9) is much lower. A possible explanation is that practitioners do not regard or use, for instance, GSN as a technique for evidence traceability.

4.4RQ4: What Techniques Are Used For Assessing Evidence In Practice?

TABLE III shows the total number of responses (N), the median, and the mode (in bold) for each evidence assessment technique. The evidence assessment techniques with the highest medians are *Checklists* and *Expert judgment in which the rationale behind the assessment is recorded*, and both techniques were reported as used by all the respondents. Therefore, these techniques seem to be the most frequently used ones in industry, with *Checklists* as the technique for which the highest ratio of respondents indicated that it is used *Always*. In contrast, *Quantitative approach* and *Expert Judgment without rationale recorded* are the only techniques with both *Never* as mode and the highest percentage of respondents indicating that they are *Never* used.

TABLE III. FREQUENCY OF USE OF EACH EVIDENCE ASSESSMENT TECHNIQUE

Evidence Assessment Technique	N	Median	Never	Rarely	Sometimes	Very Often	Always
Checklists	*51*	*Very Often*	*0% (0)*	*3.9% (2)*	***33.3% (17)***	*31.4% (16)*	*31.4% (16)*
Expert Judgment with rationale recorded	*51*	*Very Often*	*0% (0)*	*3.9% (2)*	***35.3% (18)***	***35.3% (18)***	*25.5% (13)*
Qualitative approach	*49*	*Sometimes*	*4.1% (2)*	*24.5% (12)*	*24.5% (12)*	***30.6% (15)***	*16.3% (8)*
Argumentation	*50*	*Sometimes*	*16% (8)*	*12% (6)*	*24% (12)*	***30% (15)***	*18% (9)*
Quantitative approach	*50*	*Sometimes*	***32% (16)***	*10% (5)*	*30% (15)*	*16% (8)*	*12% (6)*
Expert Judgment without rationale recorded	*49*	*Sometimes*	***26.5% (13)***	*22.4% (11)*	***26.5% (13)***	*18.4% (9)*	*6.1% (3)*

Similar to the evidence structuring techniques, some respondents mentioned additional techniques for evidence assessment. One respondent reported using techniques such as FMEA, FTA, Markov analysis, human regulators, robustness tests, and tools for coverage analysis and static analysis, DOORS, and hazard tracking databases. One respondent mentioned that evidence is assessed based on the rigor applied to produce it for (e.g., level of coverage of code).

When asked if it was checked that the confidence in a piece of evidence is related to the confidence in other pieces, and 71% of the respondents (37) acknowledged it. Similarly, 83% of the respondents (43) indicated that how a change in a piece of evidence might affect the confidence in other pieces was checked. These results provide further information about how industry deals with evidence traceability and change impact analysis, and more concretely for evidence assessment purposes. The results are also consistent with RQ2-related answers. Nonetheless, many aspects of the specific processes followed for evidence assessment remain open questions. For example, it could be studied how traceability matrices are used in the analysis of how confidence in a piece of evidence is affected by changes in other pieces.

In relation to the possibility of trying to gain further insights in the future, it might be interesting and very important to try to determine and better understand how experts decide upon and gain confidence in system safety. *Expert judgment with rationale recorded* seems to be used very often, and more knowledge about how experts judge could (1) help system suppliers record beforehand the information that a third party will require to assess safety, and thus probably reduce expenses, and (2) ideally help experts to improve their judgment. For example, ways to avoid overconfidence or other biases could be proposed if problems related to these aspects were discovered. We also wonder about the limitations and barriers that some techniques might pose, and more concretely about how practitioners address them. For example, we think that the credibility and value of expert judgement might be hindered if the rationale is not recorded. Studying the processes and techniques used in industry for deciding upon or eliciting the values for a quantitative evidence assessment would also be interesting.

In our opinion, an interesting finding corresponds to the fact the median of *Argumentation* as a technique for evidence assessment is higher than the median of *Argumentation-based graphical notations* as a technique for evidence structuring. This suggests that non-graphical means are in use for argumentation. Researchers might therefore be interested in empirically evaluating and comparing text-based and graphical argumentation.

4.5RQ5: What Challenges Do Practitioners Face Regarding Provision Of Safety Evidence?

TABLE IV shows the total number of responses (N), the median, and the mode (in bold) for each challenge in evidence provision. In this table, absence of an answer from a respondent meant that they had not faced or noticed the challenge. The median of all the challenges is *Important*. Very few respondents indicated that the challenges were *Uninmportant* or *Of Little Importance*, or that they had not faced them. The challenges reported by the highest ratio of respondents as *Very Important* were *Determination of confidence in evidence to support a particular claim about system safety*, *Compliance demonstration for systems whose compliance has not been previously demonstrated*, and *Suitability and application of safety standards*.

The reported importance of the latter challenge increases our confidence in the need for the survey. The lack of information in safety standards about how to manage evidence in practice and thus the potential problems in applying and showing compliance with them are two of the main motivations for the survey. Therefore, we consider that the results contribute to mitigating these issues. *Suitability and application of safety standards* is also the challenge for which the highest number of respondents indicated to having faced it. Although it is the challenge with the lowest number of respondents indicating that they had faced it, the importance of *Determination and decision upon the information that can be provided as evidence* also supports our claims about the relevance of the survey.

Some respondents extended the list by mentioning additional and more specific challenges. More concretely, the respondents indicated issues related to system development documentation, demonstration of compliance in a new country, tailoring certification approaches to the needs of the certification official assigned, analysing the effect of hardware on software and vice versa, and collection and maintenance of development artefacts.

TABLE IV. IMPORTANCE OF EACH CHALLENGE IN EVIDENCE PROVISION

Challenge in Evidence provision	N	Median	Unim-portant	Of little Importance	Moderately Important	Important	Very Important
Determination of confidence in evidence to support a particular claim about system safety	*48*	*Important*	*0% (0)*	*2.1% (1)*	*20.8% (10)*	***39.6% (19)***	*37.5% (18)*
Compliance demonstration for systems whose compliance has not been previously demonstrated	*48*	*Important*	*2.1% (1)*	*4.2% (2)*	*14.6% (7)*	***41.7% (20)***	*37.5% (18)*
Need for providing arguments to show how evidence meets the requirements/objectives of a safety standard	*49*	*Important*	*2% (1)*	*0% (0)*	*18.4% (9)*	***46.9% (23)***	*32.7% (16)*
Provision of adequate process information as evidence for the whole development and V&V process	*48*	*Important*	*0% (0)*	*4.2% (2)*	*18.8% (9)*	***43.8% (21)***	*33.3% (16)*
Suitability and application of safety standards	*50*	*Important*	*2% (1)*	*6% (3)*	*22% (11)*	*32% (16)*	***38% (19)***
How to effectively create and structure safety cases	*48*	*Important*	*4.2% (2)*	*4.2% (2)*	*20.8% (10)*	***35.4% (17)***	***35.4% (17)***
Compliance demonstration for new technologies	*49*	*Important*	*0% (0)*	*10.2% (5)*	*20.4% (10)*	***34.7% (17)***	***34.7% (17)***
Provision of evidence for systems that reuse existing components/subsystems	*49*	*Important*	*2% (1)*	*8.2% (4)*	*16.3% (8)*	***42.9% (21)***	*30.6% (15)*
Determination and decision upon the information that can be provided as evidence	*47*	*Important*	*0% (0)*	*6.4% (3)*	*23.4% (11)*	***44.7% (21)***	*25.5% (12)*
Existence of problems which, based on your experience, are exclusive to the application domain selected and do not arise in others	*48*	*Important*	*4.2% (2)*	*6.3% (3)*	*25% (12)*	***33.3% (16)***	*31.3% (15)*

Related studies have acknowledged the existence of similar needs and challenges. For example, previous work [21] has reported on the challenge of reusing arguments and evidence artefacts. Similarly, the challenge of suitability and application of the safety standard was discussed in [20], with respondents pointing out key issues such as the need for interpreting the standards and the complexity of understanding them. The main contribution of our survey is that it shows the perceived importance of the challenges, and the extent to which practitioners from a more general audience have faced them.

We think that it would be valuable to study why some respondents (and thus practitioners in general) have not faced or observed some challenges. For example, four respondents did not report *Determination of confidence in evidence to support a particular claim about system safety.* It might also require further investigation why and when practitioners regard some challenges as *Unimportant* or *Of Little Importance*. Evidently, provision of means for mitigating the challenges is an area that future research should address.

4.6RQ6: What Gaps Exist Between The State Of The Art And The State Of The Practice Regarding Safety Evidence Management?

In this section, we compare the results obtained from the survey with those obtained from the SLR in [7]. To represent the comparison between the practice and literature, we established a comparative scale. The scale aims to replicate the importance of the phenomena in the literature and in practice according to their frequency. The range of the scale is equally divided into three parts: *Low*, *Medium* and *High*, from the lowest to the highest frequency of the categories observed in the SLR and in practice. Although we had other ways of comparing the results (for e.g., equally splitting 100% by three ranges), in our opinion the method used is the most suitable. The two studies have unique, different sample sizes (218 publications in the SLR and 52 participants in the survey). We further believe that the comparison provides a useful overview of the current state of the art versus the state of the practice. It must also be noted that a comparison for RQ2 is not performed because such a RQ was not studied in the SLR.

For the evidence types, the scales for practice are divided equally based on the lowest frequency (17%) and highest frequency (91%) reported in the survey. Hence, the scale used is *Low* (17-41%), *Medium* (42-66%) and *High* (67-91%). Similarly, the scales for the literature are divided equally based on the lowest frequency (1%) and highest frequency (51%) observed for evidence types in the SLR. Therefore, the scale used is *Low* (1-17%), *Medium* (18-34%) and *High* (35-51%). TABLE V shows the difference in the importance given in practice and the importance observed in literature for each evidence type. The comparison shows that 16 evidence types have been given high importance in practice but observed to be of *Low* importance in literature For example, many of the testing results evidence types whose importance seems to be *High* in practice have been observed in *Low* amounts in literature. Further investigation on these differences needs to be performed in the future. Evidence types related to hazard analysis such as *Hazard specification* and *Risk analysis* results have been given equal *High* importance in both literature and practice. This might be an indication that academia has acknowledged the relevance of these evidence types and more importance has been given to them. Finally, nine evidence types have both *Low* importance in practice and literature.

Regarding evidence structuring techniques, and in line with the comparison for the evidence types, we specified the importance from the literature considering the lowest (3%) and the highest (92%) frequency of the structuring technique observed in the SLR. We then divided them equally as *Low* (3-33%), *Medium* (34%-63%), and *High* (64-92%). Likewise for evidence assessment techniques, based on the lowest (6%) and highest (68%) frequency of the assessment technique observed in the SLR, the scale was *Low* (6-26%), *Medium* (27%-47%) and *High* (48-68%). On the other hand, for the importance in practice, we used the median of a particular structuring and assessment technique as follows: *Low* (*Never/Rarely*), *Medium* (*Sometimes*), and *High* (*Very Often/Always*). *High*. TABLE VI compares the importance observed in practice and the importance observed in literature for each structuring and assessment techniques category. Three items, namely *Unstructured Text, Expert judgment without recording the rationale,* and *Expert judgment recording the rationale* were not identified in the SLR and are hence are not compared in the table.

TABLE V. COMPARISON OF IMPORTANCE GIVEN IN PRACTICE AND IMPORTANCE OBSERVED IN THE TECHNICAL LITERATURE FOR EACH EVIDENCE TYPE

Importance in practice versus Importance in the technical literature	**Evidence Types**	
High in practice vs. Low in the technical literature	• Acceptance Testing Results • Architecture Specification • Configuration Management Plan • Development Plan • Functional Testing Results • Inspection Results • Integration Testing Results • Normal Range Testing Results	• Performance Testing Results • Review Results • Safety Management Plan • System Testing Results • Test Cases Specification • Traceability Specification • Unit Testing Results • V&V Plan
High in both practice and the technical literature	• Hazards Specification	• Risk Analysis Results
Low in both practice and the technical literature	• Communication Plan • Model Checking Results • Reused Component Historical Service Data Specification	• Object Code • Robustness Testing Results • System Historical Service Data Specification • Theorem Proving Results
Medium in both practice and the technical literature	• Hazards Mitigation Specification	
Low in practice vs. Medium in the technical literature	• Accidents Specification	

Medium in practice vs. Low in the technical literature	•Activity Records •Assumptions and Conditions Specification •Automated Static Analysis Results •Development and V&V Staff Competence Specification •Modification Procedures Plan •Non-operational Testing Results •Operation Procedures Plan •Operational Testing Results •Project Monitoring Plan	•Reliability Testing Results •Reused Component Specification •Risk Management Plan •Simulation Results •Source Code •Stress Testing Results •Structural Coverage Testing Results •System Inception Specification •Tool Support Specification
High in practice vs. Medium in the technical literature	•Design Specification	•Requirements Specification
Medium in practice vs. High in the technical literature	•Hazards Causes Specification	

A stark difference in the evidence structuring techniques used in practice and the SLR is the use of *Argumentation-based graphical notations*. This technique for evidence structuring was observed the most in the SLR (*High* importance), however its frequency in practice has led to ranking its observed importance as *Low*. All the other structuring techniques have been observed in *Low* numbers in literature even though their importance in practice is either *Medium* in some cases. The results suggest that a lot of research effort has been spent on techniques that have seen little industrial adoption thus far. Researchers might therefore want to identify the reasons for this low industrial penetration by investigating possible root causes. Some possibilities are a high learning curve, the lack of adequate tool support, or a mismatch between the research and industrial needs.

When comparing the evidence assessment techniques, the main difference that we have identified is that the importance of *Checklists* in practice is *High* while in literature is *Low*. A possible reason is that the checklists used in industry correspond to well-established, widely-accepted means for evidence assessment, thus research on new checklists might not be very important. When performing the SLR, we did not consider expert judgement as a technique for evidence assessment unless the result or rationale was recorded with or based on another technique. Since the results of the survey show that the importance of this technique in practice is *High*, and as mentioned above, we think that studying how experts assess safety evidence and thus system safety is a relevant area for future research.

TABLE VI. COMPARISON OF IMPORTANCE GIVEN IN PRACTICE AND IMPORTANCE OBSERVED IN THE TECHNICAL LITERATURE FOR EACH EVIDENCE STRUCTURING AND ASSESSMENT TECHNIGUE

Importance in practice versus Importance in the technical literature	**Evidence Structuring Techniques**	
Low in practice vs. High in the technical literature	•Argumentation-based graphical notations	
Low in practice vs. Low in the technical literature	•Process models	
Medium in practice vs. Low in the technical literature	•Textual templates •Structured Text	•Conceptual/information models
Importance in practice versus Importance in the technical literature	**Evidence Assessment Techniques**	
High in practice vs. Low in the technical literature	•Checklists	
Medium in practice vs. High in the technical literature	•Qualitative approach	•Argumentation
Medium in practice vs. Low in the technical literature	•Quantitative approach	

With regards to the challenges in evidence provision and management, we ranked the importance of all the challenges in practice as *High* because their median was *Important*. Based on the lowest (7) and the highest number of publications (60) in which the challenges had been observed in the SLR, the scale was: *Low* (7-24), *Medium* (25-42) and *High* (43-60*)*. TABLE VII shows the comparison of the various

challenges in the literature and practice. Although the importance of most of the challenges is *Low* in literature, we regard as very positive that the importance of all the challenges identified in the SLR is *High*. We believe that academia is addressing the right challenges, despite weaknesses such as the low number of publications reporting on or linked to practices in industry. It is also important to mention that two challenges (*Compliance demonstration for new technologies* and *Compliance demonstration for systems whose compliance has not been previously demonstrated*) are relatively new in literature, as they were identified in publications in the last 7 years of the SLR period. These challenges have not been widely studied yet, thus it is understandable that they have ranked as of *Low* importance in the literature.

TABLE VII. COMPARISON OF IMPORTANCE GIVEN IN PRACTICE AND IMPORTANCE OBSERVED IN THE TECHNICAL LITERATURE FOR EACH CHALLENGE IN EVIDENCE PROVISION

Importance in practice versus Importance in the technical literature	**Challenges in Evidence Provision**
High in practice vs. Low in the technical literature	•Compliance demonstration for new technologies •Compliance demonstration for systems whose compliance has not been previously demonstrated •Need for providing arguments to show how evidence meets the requirements/objectives of a safety standard •Provision of evidence for systems that reuse existing components/subsystems
High in practice vs. Medium in the technical literature	•Determination of confidence in evidence to support a particular claim about system safety •Provision of adequate process information as evidence for the whole development and V&V process •Suitability and application of safety standards
High in both practice and the technical literature	•Determination and decision upon the information that can be provided as evidence •How to effectively create and structure safety cases

In general, it could be analysed and determined in the future why the differences between the state of the art and the state of the practice have been found. Such analysis might be especially relevant when some aspects have been highly reported in the literature but not by the practitioners. This could mean that practitioners have not adopted some approaches because they still need to be more mature, or that the approaches simply do not really fit industry needs. Another explanation could be unawareness of research results in industry. Aspects highly reported by practitioners but not by researchers could simply imply that industry do not face problems with these topics despite their high frequency of use. On the other hand, they could be the source of very useful new research in the case of, for instance, the challenges. In any case, and as discussed above, we think that it is essential for future research on safety evidence management to be much further evaluated in industrial settings in order to draw conclusions about its usefulness in practice.

5 CONCLUSION

This paper presented the results of a questionnaire-based survey aimed at investigating the state of the practice on safety evidence management. The results are based on 52 valid responses from 11 different domains and 15 countries. In the survey, we covered industrial perspectives and practices related to (1) the safety evidence types used, (2) the processes and means for evidence change management, (3) the evidence structuring and assessment techniques employed, and (4) the challenges that practitioners face. We further compare the state of the art and the state of practice, discussing potential improvements for future research.

The results indicate that V&V artefacts such as *V&V Plan*, *Test Results*, and *Test Case Specifications* are among the most frequently used as safety evidence, thus showing the importance of V&V for demonstrating safety. However, some verification techniques such as *Model checking* and *Theorem proving* have been reported to be used in low numbers in the industry. *Requirements Specifications* and

Design Specifications also appear to be widely used as safety evidence in the industry. Most respondents reported the use manual techniques to check evidence completeness and change impact analysis on evidence items. This suggests a lack of tool support for completeness assessment and impact analysis. Non-graphical techniques for evidence structuring such as *Textual Templates* and *Text* (*Structured* and *Unstructured)* seem to be used more often in practice than graphical notations. Investigating the impact of both graphical and text-based techniques in terms of how they facilitate communication of their intended activity could be a potential future research area. Regarding safety evidence assessment, the results suggest that *Checklists* and *Expert judgment with recorded rationale* are the most common techniques. With respect to the challenges for evidence provision, the respondents shared common perspectives and all the challenges seem to be important in practice.

When comparing the state of the art and state of the practice, the results indicate that a total of 16 evidence types have been given low importance in the literature but high in the industry, including several evidence types related to testing. Remarkable differences have been identified in the importance of *Argumentation-based graphical notations* for evidence structuring and of *Checklists* for evidence assessment. The results suggest that a lot of research effort has been spent on techniques that have thus far seen little adoption in the industry.

An overall finding is that some of the tools (e.g., DOORS) and techniques (e.g., ECO) identified by our survey are not exclusive to safety. These tools and techniques offer means for collecting and managing safety evidence, but merely applying them is not sufficient for guaranteeing safety. By identifying tools and techniques used for evidence management and also aspects for which an absence of tools and techniques is indicated, the survey provides a scope for conducting more detailed examinations. The survey however is not aimed at conducting such examinations and leaves this as an area for future work. A deeper analysis of the relationship between safety and the application of certain techniques and tools requires a careful analysis of the end-to-end usage scenarios (i.e., processes) in which the tools and techniques are applied. Such an analysis will need to address several questions, including (1) whether a given tool or technique is adequate for an intended safety-related activity or it needs to be tailored and extended, (2) when and why tools and techniques reported as frequently used are actually not applied for a specific activity (e.g., evidence change management), and (3) whether the tools and techniques in use are falling short of fulfilling the requirements for a specific activity.

We acknowledge that the results of the survey may be the opinion of only a fraction of a much larger population. Hence, we have not tried to draw strong conclusions from the results by correlating the proportional number of responses on a certain type or technique. The insights gained from the survey are nevertheless an important stepping-stone for future work and arriving at more definitive conclusions. The results can further help practitioners gain awareness of evidence management industrial practices that they could adopt or adapt, as well as of challenges that might arise.

In the future, we would like to develop automated tool support for safety evidence traceability and impact analysis. We further plan to compare the evidence types reported to the information presented in different safety standards. This will allow us to study the state of the practice of safety certification and assessment from a different perspective, and to analyse the evidence needs of specific safety standards in more depth. We would additionally like to explore the expert judgment-based evidence assessment process by devising schemes for more systematic recording of expert judgment and using the rationale for more transparent evidence assessments. Finally, another important follow-on to our current work is to analyse how practitioners perceive the importance of different tools and techniques used for safety assurance and certification, and to relate the application of these tools and techniques to the mitigation of safety risks.

ACKNOWLEDGMENTS

The research leading to this paper has received funding from the FP7 programme under the grant agreement n° 289011 (OPENCOSS), the Research Council of Norway under the project Certus SFI, and the National Research Fund of Luxembourg (FNR/P10/03 - Validation and Verification Laboratory). The authors would also like to thank the OPENCOSS partners who participated in instrument evaluation, and all the individuals who participated in the survey.

REFERENCES

[1] M. Bozzano, A. Villafiorita, Design and safety assessment of critical systems, Auerbach Pub, CRC press, 2010.
[2] A. Kornecki, J. Zalewski, Certification of software for real-time safety-critical systems: state of the art, Innovations in Systems and Software Engineering, 5 (2009) 149-161.
[3] IEC, 61508 - Functional safety of electrical / electronic / programmable electronic safety-related systems, 2005.
[4] RTCA, DO-178C - Software Considerations in Airborne Systems and Equipment Certification, 2012.
[5] CENELEC, EN 50129 - Railway applications - Safety related electronic systems for signalling, European Committee for Electrotechnical Standardisation, 2003.
[6] ISO, ISO/DIS 26262 - International Standard Road vehicles — Functional safety, 2011.
[7] S. Nair, J.L. De La Vara, M. Sabetzadeh, L. Briand, An extended systematic literature review on provision of evidence for safety certification, Information and Software Technology, 56 (2014) 689-717.
[8] S. Wilson, T.P. Kelly, J.A. McDermid, Safety case development: Current practice, future prospects, in: Safety and Reliability of Software Based Systems, Springer, 1997, pp. 135-156.
[9] M. Bouissou, F. Martin, A. Ourghanlian, Assessment of a safety-critical system including software: a Bayesian belief network for evidence sources, in IEEE Proceedings of Reliability and Maintainability Symposium, 1999, pp. 142-150.
[10] S.A. Bohner, Software change impact analysis, IEEE Computer Society Press, 1996.
[11] R.K. Panesar-Walawege, M. Sabetzadeh, L. Briand, T. Coq, Characterizing the chain of evidence for software safety cases: A conceptual model based on the IEC 61508 standard, in Third IEEE International Conference on Software Testing, Verification and Validation (ICST), 2010, pp. 335-344.
[12] M. Ivarsson, T. Gorschek, A method for evaluating rigor and industrial relevance of technology evaluations, Empirical Software Engineering, 16 (2011) 365-395.
[13] B.A. Kitchenham, S.L. Pfleeger, Personal opinion surveys, in: Guide to Advanced Empirical Software Engineering, Springer, 2008, pp. 63-92.
[14] L.-H. Eriksson, Using formal methods in a retrospective safety case, in: Computer Safety, Reliability, and Security, Springer, 2004, pp. 31-44.
[15] F. Torner, P. Ohman, Automotive Safety Case A Qualitative Case Study of Drivers, Usages, and Issues, in 11th IEEE High Assurance Systems Engineering Symposium (HASE), 2008, pp. 313-322.
[16] I. Dodd, I. Habli, Safety certification of airborne software: An empirical study, Reliability Engineering & System Safety, 98 (2012) 7-23.
[17] S. Liu, V. Stavridou, B. Dutertre, The practice of formal methods in safety-critical systems, Journal of Systems and Software, 28 (1995) 77-87.
[18] R.K. Panesar-Walawege, M. Sabetzadeh, L. Briand, Supporting the Verification of Compliance to Safety Standards via Model-Driven Engineering: Approach, Tool-Support and Empirical Validation, Information and Software Technology, 55, no. 5 (2012) 836-864.
[19] SafeCer, Deliverable D1.0.1 – State-of-practice and state-of-the-art agreed over workgroup, 2011.
[20] OPENCOSS, D4.1 - Baseline for the common certification language, 2012.
[21] OPENCOSS, D5.1 - Baseline for the compositional certification approach, 2012.
[22] OPENCOSS, D7.1 - Baseline for the process-specific needs of the OPENCOSS platform, 2012.
[23] OPENCOSS, D6.1 - Baseline for the evidence management needs, 2012.
[24] C. Robson, Real world research: A resource for social scientists and practitioner-researchers, Blackwell Oxford, 2002.
[25] S. Nair, J.L.d.l. vara, M. Sabetzadeh, D. Falessi, Management of Evidence for Compliance with Safety Standards: A Survey on the State of Practice, Technical report, Simula Research Laboratory, Norway, 2014.
[26] K.B. Wright, Researching Internet‐based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services, Journal of Computer‐Mediated Communication, 10, 2005.
[27] D. Siegle, Likert Scale, http://www.gifted.uconn.edu/siegle/research/instrument%20reliability%20and%20validity/likert.html, 2010.
[28] O. Barzilay, O. Hazzan, A. Yehudai, Using social media to study the diversity of example usage among professional developers, in Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering, ACM, 2011, pp. 472-475.
[29] T. Kanij, R. Merkel, J. Grundy, Lessons learned from conducting industry surveys in software testing, in the 1st IEEE International Workshop on Conducting Empirical Studies in Industry (CESI), 2013, pp. 63-66.
[30] R.M.d. Mello, G.H. Travassos, Would Sociable Software Engineers Observe Better?, in ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2013, pp. 279-282.
[31] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in software engineering, 2nd ed., Springer, 2012.
[32] I. Bate, T. Kelly, Architectural considerations in the certification of modular systems, Reliability Engineering & System Safety, 81 (2003) 303-324.
[33] F.J. Buckley, Implementing configuration management. Hardware, software, and firmware, IEEE Computer Society Press and Piscataway, 2nd ed., 1996
[34] D. Garwood, Bills of Material: For a Lean Enterprise, Dogwood Publishing Incorporated, 2004.
[35] C.A. Ericson, Concise Encyclopedia of System Safety: Definition of Terms and Concepts, John Wiley and Sons, 2011.