

Management of Evidence for Compliance 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 Center for Software Systems V&V, Simula Research Laboratory, P.O. Box 134, 1325 Lysaker, Norway

^b University of Luxembourg, 4 rue Alphonse Weicker, L-2721, Luxembourg

^c Fraunhofer Center for Experimental Software Engineering, 5825 University Research Ct. Suite 1300 College Park, MD 20740, USA

Abstract

In many domains such as avionics, railway, and automotive, safety-critical systems must comply with prescribed measures for safety, usually presented in the form of safety standards. Demonstrating compliance with a safety standard involves providing convincing evidence that the requirements envisaged by the standard are adequately met. Real systems are large and thus require collecting and managing large quantities of evidence throughout their lifecycle. Lack of knowledge on how to collect, structure, and assess evidence can lead to certification risks. This paper aims to provide insights into how practitioners deal with safety evidence management. Specifically, we report on a survey conducted with the goal of determining practitioners' practices and perspectives on safety evidence management. A total of 52 practitioners from 15 countries and 11 application domains responded to the survey, indicating what types of information constitute evidence for safety, how evidence is structured and assessed, how evidence evolution is addressed, and what challenges they face with regards to providing evidence. The paper further analyses the commonalities across different application domains and compares the results against the state of the art. The results notably indicate that practitioners currently undertake an extensive amount of manual work while collecting and manipulating evidence information. This can be both costly and error-prone, thus suggesting the need for better tool-support for evidence management.

Keywords Safety-critical systems, safety certification, safety assurance, safety compliance, safety standards, safety evidence, empirical study, state of the practice, survey research.

1. INTRODUCTION

Failures in safety-critical systems such as those used in the avionics, railway, and automotive domains can have catastrophic consequences [1]. To ensure that safety-critical systems cannot unduly harm people, property, or the environment, these systems are typically subject to *safety certification*, also referred to as *safety assurance*. 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].

*Corresponding author

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)

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, 5] for software in airborne systems, the CENELEC standards [6] for railway systems, and ISO26262 [7] 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*” [8]. In the context of compliance with safety standards, safety evidence is also targeted at showing fulfilment of the requirements of a standard. Some generic examples of safety evidence, among several others, are test results, system specifications, and personnel competence.

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 a 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 [9]. Safety evidence can be structured either graphically (e.g., with models) or textually.

As part of evidence management, one 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 [10]. Adequacy can be estimated qualitatively (e.g., via a confidence level) or quantitatively (e.g., via a numerical adequacy degree).

Traceability links may be 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 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 [11].

Although safety standards provide some guidance for provision of safety evidence, they are generic and often targeted at a large spectrum of systems [12]. For a specific system, practitioners may therefore have difficulties in determining what information and artefacts must be collected as evidence, how to effectively structure and assess this information, and how to capture and maintain the links between various pieces of evidence information. Furthermore, an area of growing importance in the industry is the reuse of systems and thus of evidence across different domains. This necessitates a more explicit specification of the evidence requirements in different standards, along with a mapping of the commonalities between the domains [13].

Despite the abundance of research focused on supporting and improving safety evidence management, few studies have been validated in real industrial projects or have provided empirical evidence about practices and perspectives in the industry. In a recent Systematic Literature Review (SLR) on provision of safety evidence [8], it was identified that a vast majority of the studies (73%) were not validated by means of empirical methods. Only a small fraction of the studies (14%) reflected on practices in actual projects, and even a smaller fraction (2%) surveyed practitioners’ activities and perspectives. In addition, the studies that have 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 [14]. As a result, very little knowledge exists about the state of practice on safety evidence management.

The main objective of this paper is to contribute towards addressing the above gap by providing new insights into practitioners' practices and perspective regarding safety evidence management. For this purpose, an empirical study has been conducted in the form of a questionnaire-based survey [15]. The survey was targeted at practitioners who directly participate or have 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 and previous surveys.

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 evidence structuring and for evidence assessment. 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 safety evidence management practices among different domains and the results of the survey against the state of the art.

The results of the survey can be useful both for academia 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.

The rest of the paper is organized as follows. Section II presents the related work in the area. Section III describes the research method used in our study. Section IV presents the survey results and our interpretation of the results. Section V presents a summary of the results, our conclusions, and future work. Finally, Appendix A contains tables with the comparisons among domains in terms of evidence types and challenges identified, and Appendix B shows the questionnaire designed for the survey.

2. RELATED WORK

As mentioned above, a SLR analysed the state of the art on provision of evidence for safety compliance [8] and provided a comprehensive view. This study reviewed 216 peer-reviewed papers published between 1990-2012, in order to (1) identify and classify the information and artefacts considered as evidence for safety compliance, (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 216 primary studies selected, 58 had been validated by means of some empirical method and 37 presented insights into and thus evidence about industrial practices and perspectives. These studies correspond to action research (validation in real projects by the authors themselves; 25 studies), case study (validation in real projects by practitioners different to the authors; 7 studies), or survey (validation on the basis of practitioners' perspectives; 5 studies). One paper applied both action research and survey [16]. Details of these studies can be found in [8].

When validating their work through surveys, a study reported the perspective on safety cases of ten practitioners from Swedish automotive companies [17]. Issues regarding audits of airborne software have been presented in [18]. Two studies surveyed the use of formal methods [16, 19], and one analysed the experiences and opinions concerning tool qualification according to the RTCA DO-254 guidelines [2]. In the latter survey, the authors claimed that nearly 40 complete responses had been obtained. Another recent survey related to IEC61508 can be found in [20]. The authors asked 12 practitioners

from Norway about the use of the standard and their opinion about the application of model-based techniques for supporting this activity.

Other related surveys have also been recently conducted in some European research projects. In the SafeCer project (<http://www.safecer.eu>), 19 partners completed a survey [21] and responded to questions about certification and development processes, component models, safety argumentation, and verification and validation practices. This project aims to provide support for system safety arguments and for the generation of the corresponding evidence in a compositional manner for the automotive, avionics, construction equipment, and railway domains.

The study reported in this paper has been performed in the context of OPENCROSS (<http://www.opencross-project.eu>), a project 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. Within OPENCROSS, a baseline survey was conducted concerning the state of the practice in its consortium [22-25]. 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, evaluation etc.), component reuse and modular certification, and practices involved in transparency of certification process. With regards to the evidence management practices [25], partners indicated the information included in certification document, how this information is structured and managed, and how traceability between documentation is managed.

Some work has been done in the past to compare two or more safety standards from different domains in order to identify the commonalities and differences among them [26] [13]. Nevertheless, we believe that such standard comparisons do not fully reflect the state of the practice in the domains and hence are considered out of scope for related work.

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 are usually presented at a high level of abstraction, thereby lacking sufficient detail to understand the exact viewpoints of the practitioners. 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 knowledge of the practice on safety evidence management in real-world settings. The study also has the 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 on safety evidence management.

3. RESEARCH METHOD

We conducted a survey in order to provide insights into how practitioners deal with safety evidence management. A survey is a comprehensive research method for collecting information to describe, compare, or explain knowledge and behaviour [15]. The investigation presented in this paper also corresponds to qualitative (aka 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 [27].

Based on the guidelines for survey research presented in [15], the following subsections present the research questions, the survey design, instrument evaluation, data collection, data analysis, and threats to validity.

3.1. Research Question

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

Research Questions addressed	Description
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 regarding provision of safety evidence?	The aim of this question is to identify problems that practitioners might face when having to provide safety evidence and to comply with safety standards.
RQ6. What commonalities exist among different application domains with regards to safety evidence management?	The aim of this question is to determine the similarities that exist among different application domains in terms of evidence provision and management.
RQ7. 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 [8] and our findings about the practice. Consequently, we also intend to assess past research according to industrial practices and needs.

3.2. Survey Design

We designed a cross sectional web-based survey [15], 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 (Appendix B) to collect data relevant to the RQs. In its final version, the questionnaire had 21 questions and the expected time for completing it was around 15 minutes.

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 corresponds to practitioners that directly participate or have participated in evidence management for demonstrating compliance of critical computer-based systems with safety standards. The practitioners can correspond to people that:

- 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 to collect data related to the RQs. Some parts were presented in randomized order to mitigate threats to validity of the outcome, particularly errors and omissions due to respondents' fatigue. Appendix B indicates the pages, questions, and options that were randomized. Further important highlights about the questionnaire are as follows:

- For the questions concerning the information and artefacts used as safety evidence in real project settings, a list of 49 evidence types along with a short definition for each was provided and was split into two categories namely *Process information* and *Product information*. The evidence taxonomy built as a part of the SLR reported in [8] was presented to the respondents.
- Questions were included in relation to how evidence change impact analysis is performed and how the links between various pieces of safety evidence are maintained.
- Respondents were asked to indicate the use frequency of several evidence structuring and evidence assessment techniques using a five-point frequency Likert scale adopted from [28]: *Never* (0), *Rarely* (1), *Sometimes* (2), *Very often* (3) and *Always* (4).
- Respondents were asked to rate the importance of 10 possible challenges for safety evidence provision using a five-point importance Likert scale adopted from [28]: *Unimportant* (0), *Of little Importance* (1), *Moderately Important* (2), *Important* (3) and *Very Important* (4).

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).

Finally, an optional part for participation in follow-up studies was included at the end of the questionnaire. This was the only question for which an answer was not mandatory.

3.3. Instrument Evaluation 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. 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 with five practitioners was conducted. In addition to validating the understandability of the questionnaire, this process aimed to ensure that the time required 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 the 23rd of August until the 12th of November of 2012. The survey was first advertised in several groups related to demonstration of compliance with safety standards of a social networking website for people in professional occupations. 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). After a month, during which two reminders were posted on the website, we sent a personal email invitation and subsequent reminders to some practitioners we knew. We also asked them to let colleagues know about the survey. We obtained 44 valid responses during the first month, and the remaining (eight valid responses) after the personal invitations was sent. Nonetheless, the respondents in the latter set might not be only a result of the personal invitations, but correspond to people that saw the advertisement in the website.

3.4. Subject characteristics and Data analysis

A total of 80 responses were obtained and 27 (34%) were rejected because the respondents did not answer all the required questions. Another response was rejected because the respondent did not clearly indicate the role of the organization involved. A total of 52 (65%) valid responses were thus considered for data analysis.

We obtained responses from 11 different application domains with the highest number of respondents from the *Aerospace* industry (27%), followed by the *Railway* industry (17%). Figure 1 shows the frequency in percentage of respondents from each 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 1 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 a short description about the applicability of the standard.

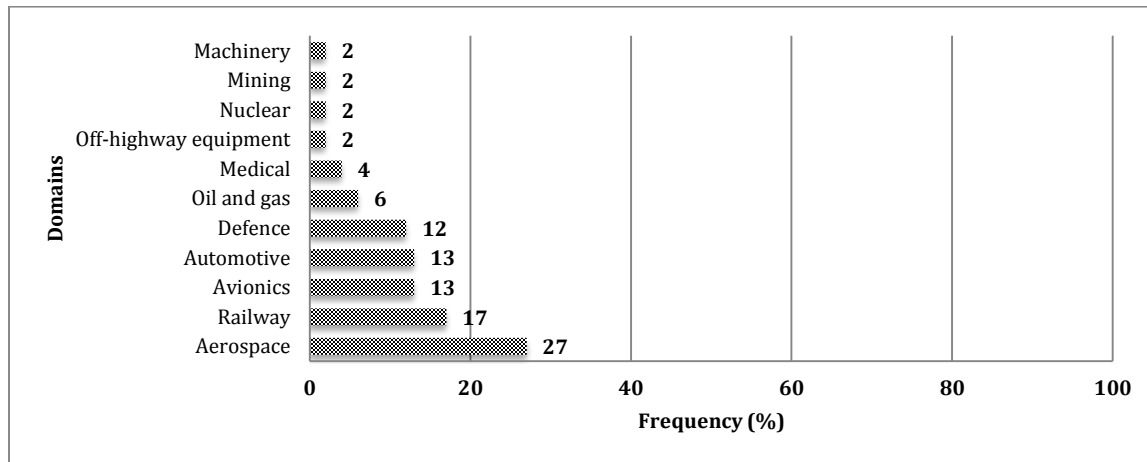


Figure 1. Frequency of response from each domain

Table 1. Safety standards identified in the survey and the percentage of respondents mentioning them

Safety Standard	Frequency	Description
RTCA DO 178B/C	33%	Standard used for software consideration of commercial and military airborne systems and equipment
CENELEC Standards	19%	Set of standards (EN50126, EN50128, and EN50129) for railway safety across Europe
IEC 61508	15%	Standard used for the certification of electrical, electronic, or programmable electronic systems
ISO 26262	13%	Standard for functional safety of road vehicles
MIL-STD-882	12%	Standard for system safety in US military
UK Def Standards 00-55/56	10%	Standard established by the Ministry of Defence (MOD) in the UK for providing safety management requirements for defence systems

RTCA DO 254	8%	Standard that provides guidance for the development of airborne electronic hardware
ARP 4754	6%	Aerospace recommendation practice for the development and certification of aircraft systems
IEC 62304	4%	Standard that specifies lifecycle requirements for the development of medical software and software within medical devices
IEC 60601	4%	Series of technical standards for the safety and effectiveness of medical electrical equipment,
ARP 4761	2%	Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment
ISO 14971	2%	Standard that establishes the requirements for risk management to determine the safety of a medical device
OHSAS 18001	2%	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%	The American Railway Engineering and Maintenance-of-way Association publishes standards and offers guidelines and best practices for railway engineering
IEC 61513	2%	Application of IEC61508 to the nuclear industry
ISO 10993	2%	A series of standards for evaluating the biocompatibility of a medical device prior to a clinical study
NORSOK	2%	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%	Standard that provides guidance on the specification, design, installation, operation and maintenance of safety instrumented functions
ISO 15998	2%	Standard that specifies performance criteria and tests for functional safety of safety-related machine-control systems using electronic components in earth-moving machinery and its equipment
JSP 454	2%	MOD Joint Service Publications that define the policy and identify specific regulatory requirements for system safety and environmental assurance for land systems.
POEMS	2%	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%	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%	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%	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%	Standard for environmental test of avionics hardware
ECSS-E-ST-40C, ECSS-E-ST-80C	2%	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%	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%	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%	It is a set of certification specifications for very light airplanes
CS-25.1309	2%	Certification specification for large airplanes
IEEE 12207	2%	Standard that establishes a common framework for software life cycle process.
Joint Software System Safety Engineers Handbook	2%	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

In relation to the country in which the respondents mainly work, we identified 15 different countries. Four respondents replied that they were involved in compliance with safety standards in several countries. Figure 2 shows the frequency in percentage of responses from each country.

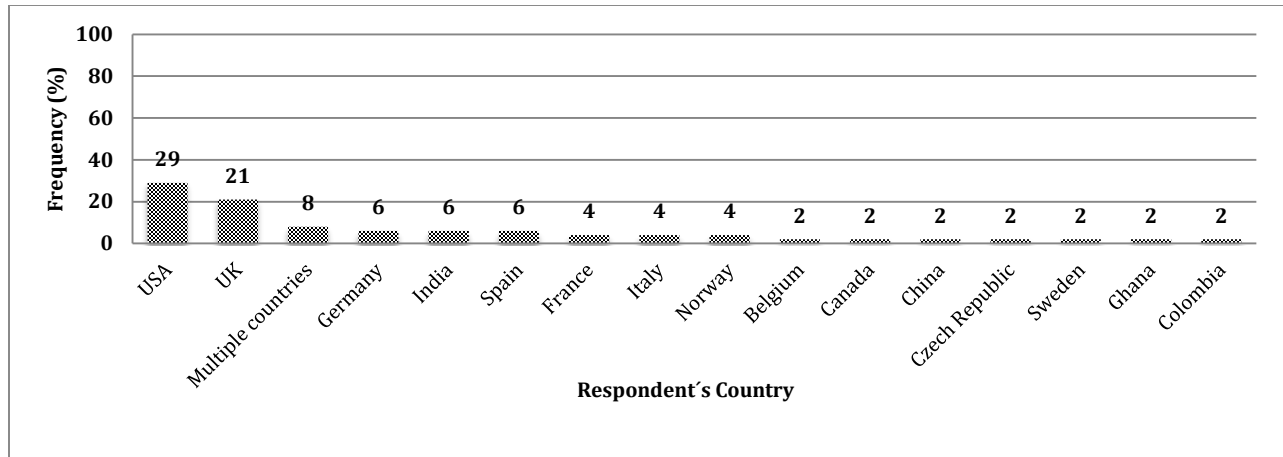
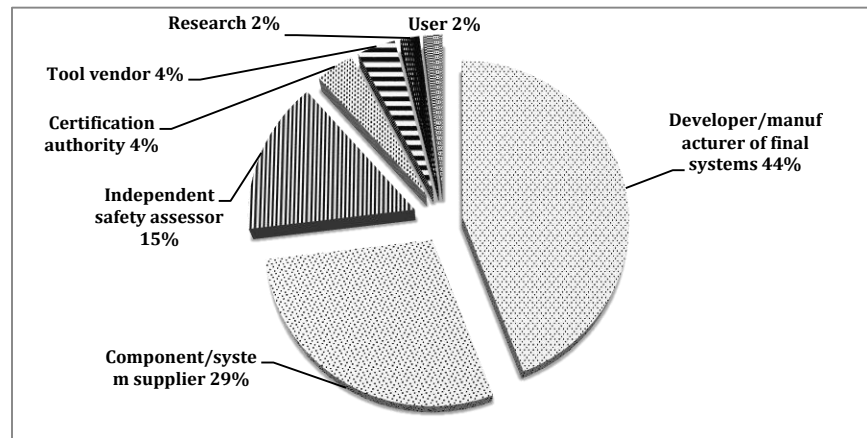
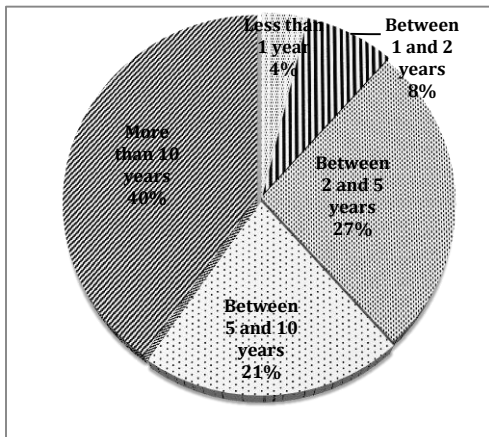


Figure 2. Frequency of response from each country

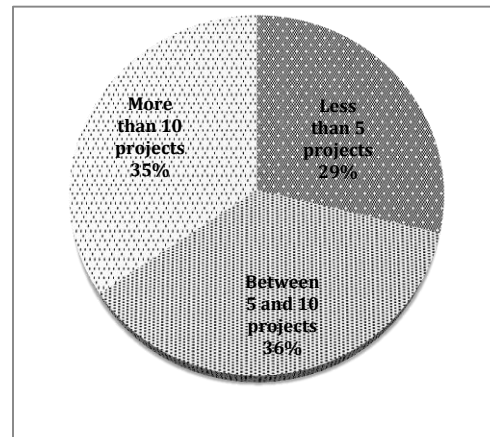
A large majority of the respondents (44%) were from *developer/manufacture of final systems* followed by *component/system supplier* (29%). Figure 3 (a) shows the percentage of respondent's organization role. About 40% of the respondents have more than 10 years of experience in demonstrating compliance with safety standards (Figure 3 (b)), and about 71% of the respondents have participated in five or more safety-critical projects (Figure 3 (c)).



(a)



(b)



(c)

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

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.

3.4. Threats to Validity

In this section, we discuss the validity threats to our study and how they were mitigated. The four perspectives presented in [29] 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. Another validity threat was the provision of options in some questions. Respondents might have found it easier to select items from a list than proposing new ones. 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 options of the questions when their understanding might be difficult or ambiguous. Instrument evaluation also mitigated this threat, and contributed to reliability of measures. 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. The background information collected for the respondents contributed to reliability of treatment implementation.

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 the potential threat of instrumentation. 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.

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.

4.1. RQ1: 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 of respondents who selected each type in the horizontal axis. *Verification & Validation plan* was the most recognized (90%) process-based evidence type. The second most selected type was *Development plan* (79%), followed by *Safety management plan* (75%) and *Configuration management plan* (71%). Only four process-based evidence types were selected by less than 50% of the respondents. They were *Operator competence specification* (27%), *Communication plan* (35%), *Reused component historical service data* (37%), and *Development and V&V staff competence specification* (46%).

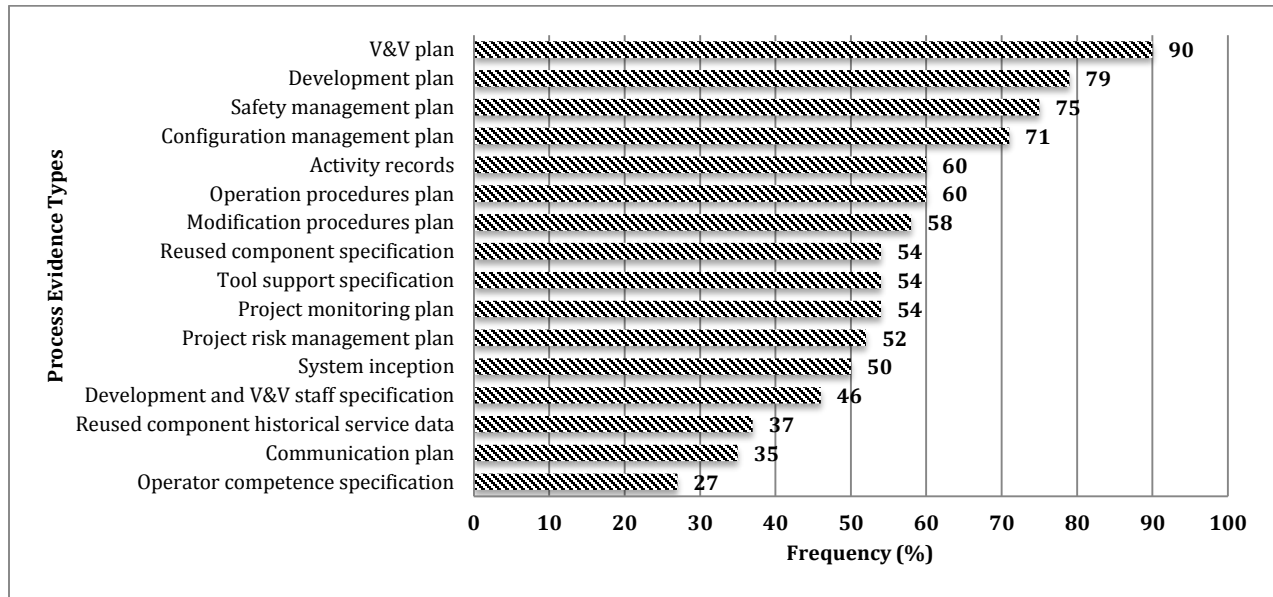


Figure 4. Frequency of each process evidence type

As for the product information category, shown in Figure 5, we identified that *Requirements specification* was the most selected product-based evidence type (87%). The second most selected type was *Test results* (85%), followed by *Test case specification* (83%) and *Design specification* (81%). The least identified evidence type in the product information was *Theorem proving results*, which was selected only 17%. Other product evidence types selected in low percentages were *Model checking* (27%), *Object code* (35%), *System historical service data* (35%), and *Accidents specification* (40%). 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 into 16 finer-grained types, shown in Figure 6. For each of these finer-grained types, we provided the respondents with a short definition obtained from our previous study [8] to help them understand the context. As indicated by percentage of respondents who selected each testing type in Figure 6, we identified that *System testing* was the most selected type in this category (89%), followed by *Functional testing* (87%), *Normal range testing* (83%), and *Acceptance testing* (81%). The least selected testing type was *Non-*

operational testing (44%). 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. This suggests high validity of construct, as the material used in the survey was complete.

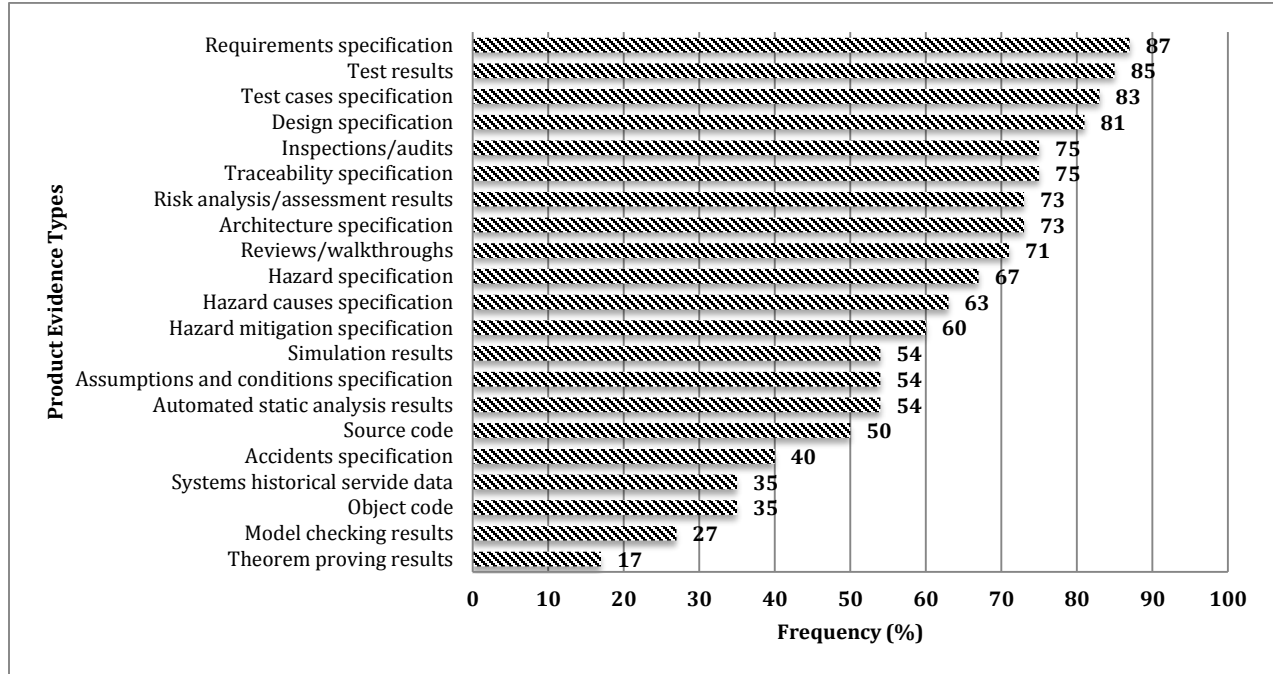


Figure 5. Frequency of each product evidence type

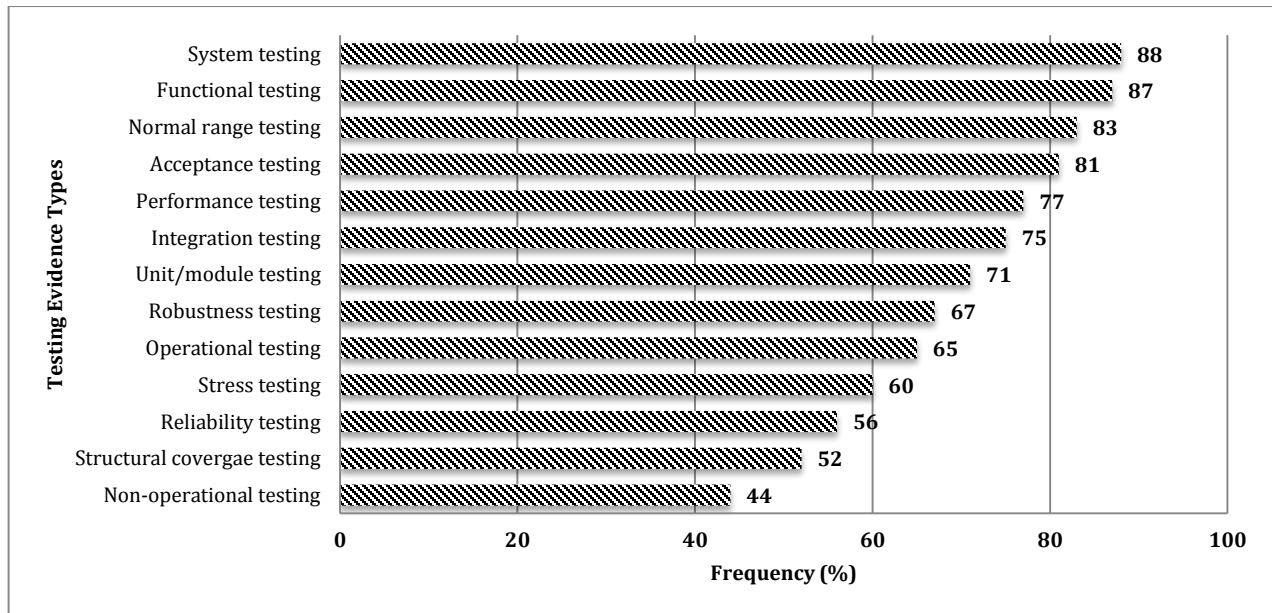


Figure 6. Frequency of each testing type

When comparing with other studies on the state of the practice, we find that the results of this survey are generally in-line with the results from previous surveys [16, 19, 25]. We consider that the differences that we have found might have been a result of the difference in scope in which the studies were

performed. For example, some papers have focused on the study of the use of formal methods, without taking other forms of safety evidence into account.

We have identified in this study that Verification and Validation related evidence types such as *Test results*, *Test case Specification*, and *V&V plan* have been among the most frequently reported evidence types. Results from the previous studies [25, 28] also show that high importance is given to the testing and verification process of a safety-critical system. Consequently, and in general, these types seem to be 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 risks and hazard 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).

4.2. RQ2: How is evidence change managed in practice?

The percentage of responses for ways to check the degree of evidence completeness is shown in Figure 7. Most of the respondents (79%) indicated that the degree of completeness for the evidence is checked manually (e.g., using paper-based checklists). Similarly, a majority of the respondents (79%) noted that they provide, check or request details about how the change of a piece of evidence has affected other pieces of evidence.

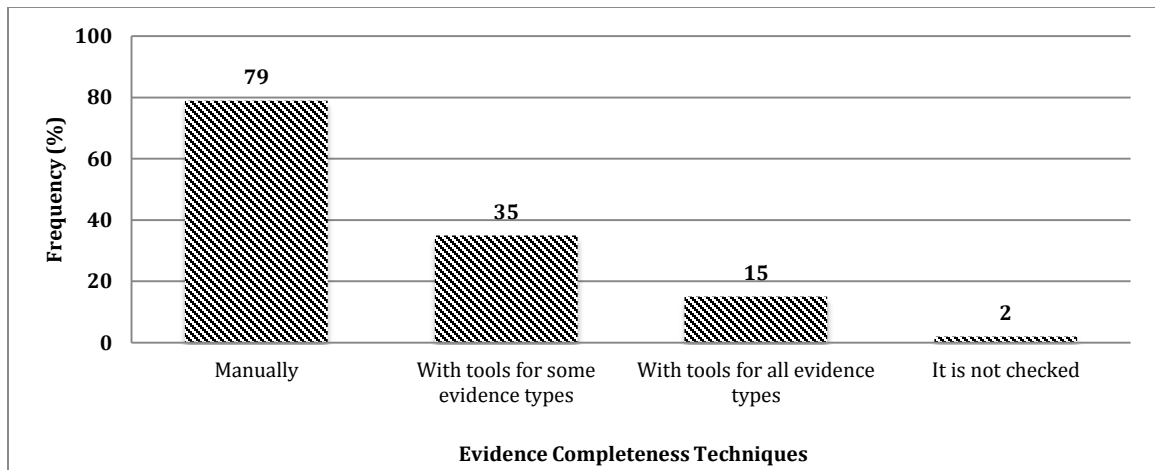


Figure 7. Frequency of techniques used for checking the degree of completeness of evidence

When asked about how they analyse the effect of the change of a piece of evidence on other pieces, 44% 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 case process [5]. Figure 8, shows the frequency of the evidence change effect techniques.

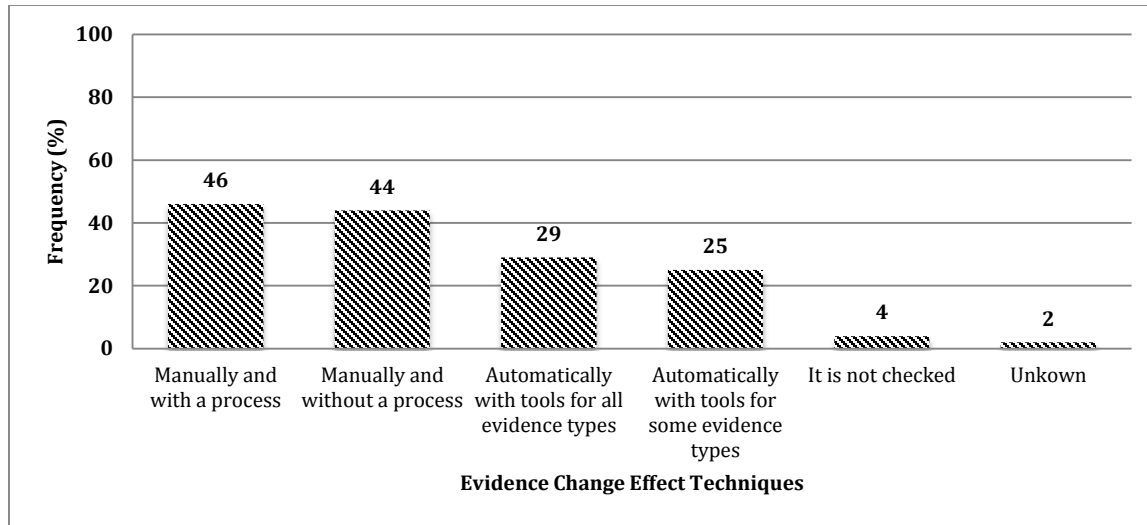


Figure 8. Frequency of techniques used for checking the effect of evidence change

The majority of the respondents (65%) replied that *Traceability matrices* are used for capturing the traceability between different pieces of evidence that they provide, check or request. 21% of the respondent indicated the use of *Models*, *Hyperlinks*, or some *Naming conventions*. Frequency of response this question is shown in Figure 9.

Some respondents provided additional information about practices for recording traceability. Single respondents acknowledged the use of:

- Engineering Change Orders (ECOs) [30]
- Bill Of Materials (BOMs) [31]
- Excel Spreadsheets
- Text documents created by version control tools and standard document templates
- Safety analysis techniques like Fault Tree Analysis (FTA), Failure Mode and Effect Analysis (FMEA), Functional Hazard Analysis and System Safety Analysis [32].

Three respondents mentioned IBM's Rational 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.

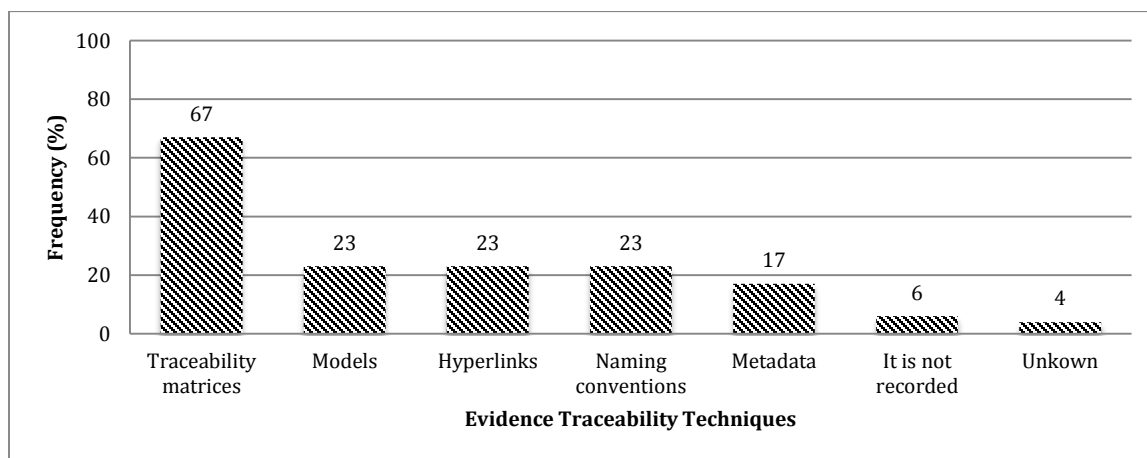


Figure 9. Frequency of each evidence traceability recording technique

When comparing the results obtained with previous surveys, we identify that the results are inline. For example, the results in [25], which was performed in a more limited setting with 15 partners from OPENCROSS, also suggest the use of traceability matrices as the most common technique for recording traceability in the safety certification documents and the results of this study report similar trends. Tools such as DOORs are also reported in previous strands of work [21] to record traceability and three respondents in this survey explicitly reported the same.

An 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 make required mitigation steps.

In our opinion, an especially relevant finding is that the results suggest that evidence change management is mainly performed manually. Given the complexity of such 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 aspect that might be studied in the future is why practitioners might not need (or find any benefit in) checking evidence completeness, analysing change impact, or recording 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 have participated) had a limited scope, or were concerned only with some specific activity, e.g., programming. Therefore, these aspects simply did not apply to them.

4.3. RQ3: What techniques are used for structuring evidence in practice?

We identified from the results that *Textual templates* were the most frequently found technique (average rating 2.29) for evidence structuring. They provide a predefined structure indicating sections to be filled for the evidence collected. The second most found structuring technique category was *Structured text* (avg. rating 2.02), which provides patterns of text in which evidence can be presented. The least found evidence structuring technique category was *Process models* such as the Software and System Process Engineering Metamodel Specification (SPEM), and *Argumentation-based graphical notation* such as the Goal Structuring Notation (GSN).

Table 2 shows the total number of response for each category, their average rating and the percentage of respondents for each Likert scale: *Never (0)*, *Rarely (1)*, *Sometimes (2)*, *Very often (3)* and *Always (4)*. The mode of the Likert scale is represented in shown in bold in the table. The average rating is calculated as follows:

$$\text{Average Rating} = \frac{(\text{Scale weight} \times \text{number of respondents for the particular scale})}{\text{Total number of respondents for the particular category}}$$

Some respondents mentioned additional techniques to structure evidence such as FTA and FMEA (1 respondent) and tools such as DOORs (2 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 RTCA DO178B/C and DO254 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.

Table 2. Average frequency rate of the evidence structuring technique categories

Evidence Structuring Technique	No. Of Respondents	Avg. Rating	"0"%	"1"%	"2"%	"3"%	"4"%
Textual templates	49	2.29	8.2	22.4	18.4	34.7	16.3
Structured Text	49	2.02	20.4	8.2	26.5	38.8	6.1
Conceptual/information models	50	1.86	18	16	36	22	8
Unstructured text	49	1.84	14.3	22.4	32.7	26.5	4.1
Argumentation-based graphical notations	49	1.45	36.7	14.3	20.4	24.5	4.1
Process models	46	1.30	32.6	30.4	17.4	13	6.5

Previous work have also acknowledged the use of textual templates documentations for structuring evidence and related aspects [20]. Another survey [22] 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. Although promising results in the use of models for structuring and managing evidence have been reported in [20], it seems that such approaches are not very often used in industry yet. The scope of the related work (in terms of the countries from which the respondents are) might be a possible explanation for such differences.

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.

4.4. RQ4: What techniques are used for assessing evidence in practice?

Table 3, shows the number of responses for each category of techniques, their average rating, and the percentage of respondents for each Likert scale (0-4). The mode of the scale is shown in bold.

The most frequently reported evidence assessment technique category was *Checklists* (average rating 2.90) closely followed by using *Expert judgment in which the rationale behind the assessment is recorded* (avg. rating 2.82). The least reported category in terms of frequency was *Expert judgment in which the rationale behind the assessment was not recorded*. Quantitative approaches such as Bayesian Belief Networks [10] were the second least frequently reported.

Similar to the evidence structuring techniques, some respondents mentioned additional techniques for evidence assessment. For example, 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 them for (e.g., level of coverage of code).

Table 3. Average frequency rate of the evidence assessment technique categories

Evidence Assessment Technique	No. Of Respondents	Avg. Rating	"0"%	"1"%	"2"%	"3"%	"4"%
Checklists	51	2.90	0	3.9	33.3	31.4	31.4
Expert Judgment with rationale recorded	51	2.82	0	3.9	35.3	35.3	25.5
Qualitative approach	49	2.31	4.1	24.5	24.5	30.6	16.3
Argumentation	50	2.22	16	12	24	30	18
Quantitative approach	50	1.66	32	10	30	16	12
Expert Judgment without rationale recorded	49	1.55	26.5	22.4	26.5	18.4	6.1

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 selected “Yes”. The remaining 29% mentioned “No”. Similarly, 83% of the respondents mentioned indicated that how a change in a piece of evidence might affect the confidence in other pieces was checked, and 17% replied “No”.

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.

In our opinion, an interesting finding corresponds to the fact the average rating of *Argumentation* as a technique for evidence assessment is higher than the rating 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.

4.5. RQ5: What challenges do practitioners face regarding provision of safety evidence?

The most highly rated challenge in terms of importance by the practitioners was determining the confidence in evidence to support a particular claim about system safety. There were two categories that were the second most highly rated in terms of their importance namely demonstrating compliance for new type of systems which has not been previously demonstrated (e.g., a legacy systems) and the need for providing argumentation to show how the evidence satisfy the safety standards requirements. The least rated challenge in terms of importance was the existence of problems exclusive to a particular application domain. The second least rated challenge was in determining the information that can be provided as evidence.

Although some challenges were rated more important than others, the difference in their importance were low, with the maximum difference of the average rating between the highest ranked and lowest ranked challenge being only 0.32. Table 4 shows the number of responses for each category, their average rating and the percentage of respondents for each Likert scale (0-4, from *Unimportant* to *Very Important*). As seen in the table, not all the respondents selected all the challenges. Absence of an answer from a respondent would mean that they had not faced or noticed the challenge.

Some respondents extended the list of options provided by mentioning additional and more specific challenges:

- *Issues relating to documentation* – two respondents from the *Avionics* domain reported the challenge of creating and maintaining documents of the various activities in the development of a critical system and the documentation of some critical safety parts and subsystems.
- *Demonstration of compliance in a new country* – one respondent from the *Railway* domain reported the challenge of demonstrating compliance for a system used in a country whose compliance has been already demonstrated for another country.
- *Tailoring certification approaches to the needs of the certification official assigned* – two respondents from the *Avionics* domain noted the challenge of modifying their safety assurance and certification process to meet the additional demands and evidence requirements of the regulatory personnel.
- *Analysing the effect of hardware on software and vice versa* – one respondent from the *Railway* domain mentioned that it is very important to assess the effect of hardware on the software and the effect of software on hardware.

- *Collection and maintenance of development artefacts* – one respondent from the *Avionics* domain stated that gathering and maintaining the development artefacts along with the decision process involved to collect them is a critical challenge. The respondent mentioned that the structure and presentation of the data collected could always be changed or corrected if all the required artefacts are collected beforehand.

Table 4. Average importance rate of the challenges in evidence provision

Challenges in Evidence provision	No. Of Respondents	Avg. Rating	“0”%	“1”%	“2”%	“3”%	“4”%
Determination of confidence in evidence to support a particular claim about system safety	48	3.13	0	2.1	20.8	39.6	37.5
Compliance demonstration for systems whose compliance has not been previously demonstrated	48	3.08	2.1	4.2	14.6	41.7	37.5
Need for providing arguments to show how evidence meets the requirements/objectives of a safety standard	49	3.08	2	0	18.4	46.9	32.7
Provision of adequate process information as evidence for the whole development and V&V process	48	3.06	0	4.2	18.8	43.8	33.3
Suitability and application of safety standards	50	2.98	2	6	22	32	38
How to effectively create and structure safety cases	48	2.94	4.2	4.2	20.8	35.4	35.4
Compliance demonstration for new technologies	49	2.94	0	10.2	20.4	34.7	34.7
Provision of evidence for systems that reuse existing components/subsystems	49	2.92	2	8.2	16.3	42.9	30.6
Determination and decision upon the information that can be provided as evidence	47	2.89	0	6.4	23.4	44.7	25.5
Existence of problems which, based on your experience, are exclusive to the application domain selected and do not arise in others	48	2.81	4.2	6.3	25	33.3	31.3

Although not explicitly, related work [22-25] have also acknowledged similar needs and challenges in the context of evidence provision and management for certification.

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*, in spite of being the challenge with the highest average rating. It might also require further investigation why and when practitioners regard some challenges as unimportant. For example, *How to effectively create and structure safety cases* is the challenge that has been most frequently reported as unimportant. In line with the discussions above about aspects for future research related to other RQs, the results obtained might have been due to the specific characteristics of the respondents.

4.6. RQ6: What commonalities exist among different application domains with regards to safety evidence management?

In this section, we compare the results obtained for RQ1-5 among the 11 domains of the respondents. The comparison is based on the identification of the evidence types, structuring and assessment techniques, and the challenges concerning evidence provision in each domain. This allows us to determine the commonalities in the domains.

Six domains namely *Oil & Gas*, *Medical*, *Off-highway Equipment*, *Nuclear*, *Mining*, and *Machinery* are not considered for discussion. We consider that the number of respondents (data points) from these domains is too low to make any conclusive remark. Nonetheless, the results obtained from these domains are shown in a separate table in Appendix A. The domains used to discuss the commonalties are *Aerospace*, *Railway*, *Avionics*, *Automotive*, and *Defence*.

Five evidence types have been reported by more than 90% of the respondents in the five domains considered for discussion. These types are *Functional Testing Results*, *Requirements Specification*, *System Testing Results*, *Test Cases Specification*, and *V&V Plan*. Another four evidence types were reported 80% or more. These are *Acceptance Testing Results*, *Design Specification*, *Development Plan* and *Inspection Results*. Table 5, shows the percentage of respondents for each evidence type in the five domains considered for discussion. The evidence types that are reported by all the respondents in each domain are shown in bold. The percentages of respondents for the remaining domains are shown in Table A-1 (Appendix A).

Table 5. Percentage of respondents for each evidence type in the five application domains discussed

Evidence Types	Aerospace %	Railway %	Avionics %	Automotive %	Defence %
Acceptance Testing Results	86	100	71	43	100
Accidents Specification	36	56	14	43	83
Activity Records	50	89	57	43	83
Architecture Specification	79	67	71	71	67
Assumptions and Conditions Specification	29	67	43	86	83
Automated Static Analysis Results	50	56	71	57	50
Communication Plan	29	22	71	29	33
Configuration Management Plan	79	89	86	43	83
Design Specification	86	100	57	71	83
Development and V&V Staff Competence Specification	43	78	57	29	50
Development Plan	86	78	100	71	83
Functional Testing Results	86	89	86	86	100
Hazards Causes Specification	50	89	43	71	83
Hazards Mitigation Specification	50	67	43	57	83
Hazards Specification	64	89	29	71	83
Inspection Results	79	78	100	71	83
Integration Testing Results	72	78	86	57	100
Model Checking Results	29	22	14	29	50
Modification Procedures Plan	43	67	57	71	83
Non-operational Testing Results	43	44	71	0	67
Normal Range Testing Results	93	89	86	57	67
Object Code	36	22	71	14	33
Operation Procedures Plan	43	67	71	43	83
Operational Testing Results	64	89	71	57	67
Operator Competence Specification	7	56	14	0	67
Performance Testing Results	71	89	57	57	100
Project Monitoring Plan	36	56	71	71	67
Reliability Testing Results	36	67	29	43	100
Requirements Specification	93	100	100	71	83
Reused Component Historical Service Data Specification	36	44	43	29	33
Reused Component Specification	50	67	43	43	67
Review Results	79	67	100	57	67
Risk Analysis Results	64	100	43	71	83
Risk Management Plan	64	56	43	14	67
Robustness Testing Results	71	67	71	57	100
Safety Management Plan	64	100	71	71	83
Simulation Results	43	67	57	43	67
Source Code	50	22	86	57	33
Stress Testing Results	71	67	71	43	67
Structural Coverage Testing Results	71	33	57	57	67
System Historical Service Data Specification	21	67	43	29	33

System Inception Specification	57	67	43	43	33
System Testing Results	100	100	71	86	100
Test Cases Specification	86	100	100	86	83
Theorem Proving Results	14	33	29	0	17
Tool Support Specification	64	56	57	43	67
Traceability Specification	93	78	86	71	50
Unit Testing Results	72	78	86	57	83
V&V Plan	93	100	100	100	83

Although safety standards such as ISO26262 for automotive domain recommend the use of formal verification for verifying the software design and implementation, none of the respondents from the automotive domain reported the use of *Theorem Proving Results* as evidence. Formal verification results such as *Theorem Proving Result* and *Model Checking Results* were reported less than 30% in total by all the domains as formal verification evidence types. Likewise, none of the respondents from the automotive domain reported the use of *Operator competence specification* as an evidence type. This can be a result of the fact that the capability of the driver does not qualify as a safety threat during automotive system certification.

Concerning evidence change management, more than 50% of the respondents from each of the five domains reported that for the evidence that they provide, check or request, the degree of completeness for the evidence is checked *Manually*. In two domains, *Railway* and *Avionics*, all the respondents indicated the use of manual methods to assess evidence completeness along with small percentage of other techniques. In domains such as *Automotive* and *Defence*, none of the respondents reported the use of tools to perform this task for all the evidence types. A rate of 11% of the respondents from *Railway*, 29% from the *Aerospace*, and 43% from *Avionics* reported the use of tools that store and provide information about the degree of completeness for all types of evidence.

Similarly, more than 40% of the respondents from each of the five domains have reported that they check the effect on other pieces of evidence due to a change in a piece of evidence through *manual predefined processes*. Such manual processes can be very time consuming when having to check large quantities of information and can be complex and error-prone. The results once again suggest the strong need for tool support to perform change impact analysis on safety evidence.

More than 50% of the respondents in all the five domains have reported the use of *Traceability Matrices* to record traces between different pieces of evidence. For the *Railway* domain, 100% of respondents reported the use of matrices and none reported the use of *Metadata* for traceability. Similarly none of the respondents from the *Automotive* domain reported the use of *Hyperlinks* to maintain traces. Table 6 shows the percentage of respondents from each domain for each of the options provided for recording traceability. Table A-2 (Appendix A), shows those domains that weren't part of the comparison due to low responses.

Table 6. Percentage of respondents for each evidence traceability recording technique in the five application domains discussed

Domain / (No. of respondents)	Evidence Traceability Techniques					
	Traceability Matrices	Models	Metadata	Hyperlinks	Naming conventions	Not recorded
Aerospace (14)	79	14	29	21	21	0
Railway (9)	100	22	0	33	22	0
Avionics (7)	57	14	29	29	43	0
Automotive (7)	71	43	29	0	14	0
Defence (6)	67	33	17	33	33	17

Table 7 is split into two cells for each category of evidence structuring techniques in every domain, except for the domains that had low number of respondents. The left cell in each category shows the

frequency of the mode. The right cell in each category is the average rating for each technique in that domain. For example, regarding the frequency of *Process Models* (sixth column) in the *Avionics* domain (fourth row), 29% of the respondents reported either never, rarely or sometimes used, checked, or requested it in practice. The average rating of *Process Models* in the *Avionics* domain was 1.29. The number of respondents for each domain is shown in brackets next to the domain name. Table A-3 (Appendix A) shows the rest of the domains that are not used for discussion.

In domains such as *Avionics* and *Automotive*, more than half of the respondents reported *Structured Text* as very often used. On the contrary, in the *Aerospace* domain 46% of the respondents reported they never use *Structured Text* technique for evidence structuring. Although there seems to be a difference in the way *Structured Text* is used among these domains, the average rating in the five domains remains 2.13 (Sometimes), with less than 1 point of difference with the average in each domain. Similarly, 46% and 43% of the respondents from the *Aerospace* and *Avionics* domains, respectively, have reported the use of *Textual Templates* as always used and 71% of the respondents from the *Automotive* domain have reported it as very often, the average rating in total is only 2.36. This suggests that some differences exist, in the sense that some practitioners highly use some techniques whereas others do not.

Table 7. Average rating and mode of the scale for each evidence structuring technique in the five application domains discussed

Domain / (No. of respondents)	Evidence Structuring Techniques											
	Structured Text		Textual Template		Argumentation-Based Graphical Notation		Conceptual/Information Models		Process Models		Unstructured Text	
Aerospace (14)	46% N	1.54	46% A	2.00	54% N	1.00	54% N	1.71	55% N	0.91	33% S	1.83
Railway (9)	44% VO	1.89	44% S	2.00	33% VO	2.00	44% S	2.33	44% N	1.00	33% VO	2.11
Avionics (7)	57% VO	2.86	43% A	3.14	43% N	1.29	29% R/S/VO	2.29	29% N/R/S	1.29	43% S	1.43
Automotive (7)	57% VO	2.57	71% VO	3.29	57% S	1.57	29% N/S/VO	1.57	33% R	1.83	57% R	1.57
Defence (6)	80% S	1.80	60% R	1.40	33% VO	2.17	33% R/S	1.50	67% S	1.67	50% S	2.17

*A- Always; VO- Very Often; S- Sometimes; R- Rarely; N- Never

Argumentation-Based Graphical Notations and *Conceptual/information models* were reported as *Sometimes* used on average, in all the five domains. However, we observed differences among the responses inside a given domain on how frequently these techniques are used. For example, 29% of the respondents from the *Avionics* and *Automotive* domain have reported different scales (*never*, *rarely* or *very often*) of using *Conceptual models*, showing different practices within the same domains.

When comparing the domains in terms of how they assess evidence adequacy, the difference in the average rating of *Expert judgment without recording the rationale* between the *Railway* domain and the *Avionics* domain is more than 1. Similarly the difference between *Railway* and *Automotive* is more than 1, suggesting considerable difference in how the technique is used among the domains. All the five domains have acknowledged the use of *Expert judgment with rationale recorded* as very often used, thus showing similarities in these domains. Difference can be found in the use of *Quantitative Methods* to assess evidence between *Defence* and *Railway*. The average rating of this technique in *Railway* is 2.56 while in *Defence* is only 0.83. Differences exist among the automotive responses, as 43% of the respondents have reported never for *Quantitative Methods* while another 43% have reported very often. Once again, the results indicate likely differences in practices within individual domains. The difference in the average rating of *Qualitative Methods* between *Railway* and *Avionics* is 1.65, suggesting a

considerable difference in using this technique between the two domains. All the domains have acknowledged *Checklists* as very often used, with *Defence* reporting the lowest average rating (2.33) and *Avionics* reporting the highest (3.71). Table 8, shows the mode of the scale for a given technique and the average rating for each assessment technique. Table A-4 (Appendix A) shows the information for the remaining domains.

Table 8. Average rating and mode of the scale for each Evidence Assessment Techniques in all the five application domains discussed

Domain / (No. of respondents)	Evidence Assessment Technique											
	Expert judgment without rationale		Expert judgment with rationale		Argumentation		Quantitatively		Qualitatively		Checklists	
Aerospace (14)	31% N/R	1.46	43% A	2.86	31% N/VO	1.85	33% N	1.50	50% VO	2.50	46% VO	3.08
Railway (9)	44% VO	2.22	67% VO	3.33	44% VO	2.78	44% S	2.56	56% A	3.22	44% S	2.89
Avionics (7)	33% N/R/S	1.00	43% VO	3.00	67% S	2.00	43% N	1.43	29% N/S/VO	1.57	71% A	3.71
Automotive (7)	57% S	1.14	50% S	2.67	43% S	2.43	43% N/VO	1.86	40% S	2.40	43% S/VO	2.71
Defence (6)	33% R/S	1.50	50% S/VO	2.50	33% S/A	2.50	50% N	0.83	50% R	1.67	50% S	2.33

*A- Always; VO- Very Often; S- Sometimes; R- Rarely; N- Never

With regards to the challenges in evidence provision, all the challenges have been either reported as moderately important, important, or very important in all the domains (both used for discussion and those which are not part of the discussion). The average rating in total of all the challenges in the five domains is almost 3. This indicates that all the domains seem to be facing similar obstacles when having to provide safety evidence for compliance with safety standards. Table 9, shows the comparison of the challenges among the five domains used for discussion. The comparison of the remaining six domains can be found in Table A-5 (Appendix A).

Table 9. Average rating and mode of the scale for each evidence provision challenges faced in the five application domains discussed

Challenges	Domain / (No. of respondents)									
	Aerospace (14)		Railway (9)		Avionics (7)		Automotive (7)		Defence (6)	
<i>Compliance demonstration for new technologies</i>	43% I	2.71	50% I/VI	3.5	57% VI	3	57% I	2.86	60% VI	3.4
<i>Compliance demonstration for systems whose compliance has not been previously demonstrated</i>	36% I/VI	2.93	75% VI	3.63	57% I	3.14	67% I	3.33	40% MI/I	2.8
<i>Determination and decision upon the information that can be provided as evidence</i>	50% I	2.93	38% I	2.75	43% VI	2.86	83% I	2.83	50% VI	3.17
<i>Determination of confidence in evidence to support a particular claim about system safety</i>	43% I	2.93	50% I	3.25	43% MI/VI	3	50% VI	3.33	40% I/VI	3.2
<i>Existence of problems which, based on your experience, are exclusive to the application domain selected and do not arise in others</i>	36% VI	2.79	38% MI/I	2.88	57% VI	3	43% VI	3.14	60% I	2.4
<i>How to effectively create and structure safety cases</i>	36% VI	2.43	56% VI	3.44	43% VI	2.86	50% I	3.17	40% MI/VI	3
<i>Need for providing arguments to show how evidence meets the requirements/objectives of a safety standard</i>	50% I	3.07	75% I	3.25	43% I/VI	3	50% I	3.17	60% MI	2.4

<i>Provision of adequate process information as evidence for the whole development and V&V process</i>	36% MI	2.79	50% VI	3.38	71% I	3.57	71% I	2.86	80% I	2.8
<i>Provision of evidence for systems that reuse existing components/subsystems</i>	36% I	2.64	50% I	3.25	71% I	3.29	33% I/VI	2.83	60% VI	3.2
<i>Suitability and application of safety standards</i>	36% I/VI	2.86	50% VI	3.38	57% VI	3	43% VI	3.14	50% I	2.83

*VI- Very Important; I- Important; MI-Moderately Important; OI-Of little Importance; UI- Unimportant

Given the interest in industry in the correspondence of safety assurance and certification practices and needs among different domains, it could be interesting to study if some variations would be found in relation to this paper in surveys conducted in the future, assuming that such cross-domain possibilities are finally realized in industry. Other aspects that might be further studied are (1) to what extent the differences between domains can hinder reuse among them, (2) why some techniques for evidence structuring and for evidence assessment seem to be more frequently used in some domains than others, and (3) to what extent practitioners in a given domain can benefit from the adoption of safety evidence management practices commonly used in other domains.

4.7. RQ7: 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 this study with those obtained from the SLR in [8]. To represent the comparison between the practice and literature for evidence types, we establish comparative scale. The scale is established to replicate the importance given to the types in the literature and in practice according to the 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 3 ranges), in our opinion, the method used in this paper seems to be the most optimal given that the two studies have unique sample size (216 studies in the SLR and 52 participants in the survey). Nonetheless we believe that the comparison provides a useful overview of the current state of the art versus the state of the practice.

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 (52%) observed for evidence types in SLR. Therefore, the scale used is *Low* (1-17%), *Medium* (18-34%) and *High* (35-52%). Table 10 uses three different shades to show the difference in literature and practice for each evidence type. The darker shade shows the large contrast (*High* vs. *Low*), the lighter shades shows low contrast (*Medium* vs. *Low* or *High* vs. *Medium*) and no shade show same level of importance. The comparison shows that a number of evidence types that have been given high importance in practice have been observed to be of low or medium importance in literature. Only 11 evidence types have had the same level of importance given equally both in literature and in practice. 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 types of evidences and more importance has been given to them. On the other hand, many of the testing results evidence types reported as *high* in practice have been observed in *low* amounts in literature.

To represent the comparison between the practice and literature for evidence structuring techniques, we use the following three-level comparative scale based on the lowest (4%) and the highest (91%) frequency of the categories in SLR: *Low* (4-33%), *Medium* (34%-63%) and *High* (64-91%). For

evidence assessment techniques, based on the lowest (6%) and highest (67%) frequency observed in SLR, the scale was equally divided into: *Low* (6-26%), *Medium* (27%-47%) and *High* (48-67%). On the other hand, for the scale of the practice, we map the average rating of a particular structuring and assessment techniques as follows: (0-1.33) -> *Low*, (1.34-2.66) -> *Medium* and (2.67-4) -> *High*. Table 11 uses the same three different shades as the above to compare the structuring and assessment techniques. 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 marked as *Not Applicable (NA)* in the table.

Table 10. Comparison of importance given practice and importance observed in literature for each Evidence Types

Evidence Types	Importance given in practice	Importance observed in literature
Acceptance Testing Results	High	Low
Accidents Specification	Low	Medium
Activity Records	Medium	Low
Architecture Specification	High	Low
Assumptions and Conditions Specification	Medium	Low
Automated Static Analysis Results	Medium	Low
Communication Plan	Low	Low
Configuration Management Plan	High	Low
Design Specification	High	Medium
Development and V&V Staff Competence Specification	Medium	Low
Development Plan	High	Low
Functional Testing Results	High	Low
Hazards Causes Specification	Medium	High
Hazards Mitigation Specification	Medium	Medium
Hazards Specification	High	High
Inspection Results	High	Low
Integration Testing Results	High	Low
Model Checking Results	Low	Low
Modification Procedures Plan	Medium	Low
Non-operational Testing Results	Medium	Low
Normal Range Testing Results	High	Low
Object Code	Low	Low
Operation Procedures Plan	Medium	Low
Operational Testing Results	Medium	Low
Operator Competence Specification	Low	Low
Performance Testing Results	High	Low
Project Monitoring Plan	Medium	Low
Reliability Testing Results	Medium	Low
Requirements Specification	High	Medium
Reused Component Historical Service Data Specification	Low	Low
Reused Component Specification	Medium	Low
Review Results	High	Low
Risk Analysis Results	High	High
Risk Management Plan	Medium	Low
Robustness Testing Results	Low	Low
Safety Management Plan	High	Low
Simulation Results	Medium	Low
Source Code	Medium	Low
Stress Testing Results	Medium	Low
Structural Coverage Testing Results	Medium	Low
System Historical Service Data Specification	Low	Low
System Inception Specification	Medium	Low
System Testing Results	High	Low
Test Cases Specification	High	Low

Theorem Proving Results	<i>Low</i>	<i>Low</i>
Tool Support Specification	<i>Medium</i>	<i>Low</i>
Traceability Specification	<i>High</i>	<i>Low</i>
Unit Testing Results	<i>High</i>	<i>Low</i>
V&V Plan	<i>High</i>	<i>Low</i>

Table 11. Comparison of average frequency in practice and importance given in literature for each Evidence Structuring and Assessment techniques

Evidence Structuring Techniques	Average frequency of use in practice	Importance observed in literature
Textual templates	<i>Medium</i>	<i>Low</i>
Structured Text	<i>Medium</i>	<i>Low</i>
Conceptual/information models	<i>Medium</i>	<i>Low</i>
Unstructured text	<i>Medium</i>	<i>NA*</i>
Argumentation-based graphical notations	<i>Low</i>	<i>High</i>
Process models	<i>Medium</i>	<i>Low</i>
Evidence Assessment Techniques		
Checklists	<i>High</i>	<i>Low</i>
Expert Judgment with rational recorded	<i>High</i>	<i>NA*</i>
Qualitative approach	<i>Medium</i>	<i>High</i>
Argumentation	<i>Medium</i>	<i>High</i>
Quantitative approach	<i>Medium</i>	<i>Low</i>
Expert Judgment without rational recorded	<i>Medium</i>	<i>NA*</i>

*Not applicable as the SLR did not collect these results

A stark difference in the evidence structuring techniques used in practice and SLR is the use of *Argumentation-based graphical notations*. This technique for evidence structuring was observed the most in the SLR, however its frequency in the practice has been reported to be low. All the other structuring techniques have been observed in low numbers even though their frequencies of use in practice are medium. The results suggest that a lot of research effort has been spent on a technique that has seen little industrial adoption thus far. Researchers might therefore want to identify the reasons for this low industrial penetration by investigating possible root causes such as lack of technology transfer, high learning curve, lack of adequate tool support, or a mismatch between the research and industrial needs. Research may further need to expand to cover other techniques that the practitioners more frequently use.

When comparing the evidence assessment techniques, we identified a substantial difference in what has been researched and what is being used in practice. Respondents have reported the frequency of use for *Checklists* as *High*, while in the SLR we identified only 17% of the studies addressing this technique, putting the technique into the *Low* category in terms of the level of research. However other assessment techniques such as *Qualitative assessment* and *Argumentation* have been observed to have low contrast in both literature and in practice. When performing the SLR, we did not consider expert judgement as a technique for evidence assessment, but the results of the practice show that there is a high frequency of such techniques used in practice. This might be a potential area of future research.

With regards to the challenges in evidence provision and management, and as discussed above (Section IV.4.5), all the mentioned challenges in the survey had an average importance scale of *Important*. Therefore, we mapped the average frequency of *Important* -> *High*. To compare the results, based on the lowest number of studies (7) and highest number of studies (60) observed in SLR, the scale was equally divided into: *Low* (7-25), *Medium* (26-44) and *High* (45-60). Table 12, shows the comparison of the various challenges in the literature and practice. Only two challenges namely, *Determination and decision upon the information that can be provided as evidence* and *How to effectively create and structure safety cases* have been given equal importance. On the contrary, the

challenges *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* and *Suitability and application of safety standards* have been observed in low numbers in the SLR. These challenges have been reported as important from the practitioner's point of view as the others.

In general, it could be analysed and determined in the future why any potential difference between the state of the art and the state of the practice has 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 they simply do not really fit their 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 for very useful new research, for instance, in the case of the challenges. In any case, we think that it is necessary to try and promote technology transfer from academia to industry, rather than simply proposing new approaches. As discussed above, it is essential for research on safety evidence management to be evaluated in industrial settings in order to draw conclusions about its usefulness in practice.

Table 12. Comparison of average importance in practice and the importance given in literature for each challenges in evidence provision

Challenges in Evidence Management	Average importance of use in practice	Importance observed in literature
Compliance demonstration for new technologies	High	Low
Compliance demonstration for systems whose compliance has not been previously demonstrated	High	Low
Determination and decision upon the information that can be provided as evidence	High	High
Determination of confidence in evidence to support a particular claim about system safety	High	Medium
Existence of problems which, based on your experience, are exclusive to the application domain selected and do not arise in others	High	NA*
How to effectively create and structure safety cases	High	High
Need for providing arguments to show how evidence meets the requirements/objectives of a safety standard	High	Low
Provision of adequate process information as evidence for the whole development and V&V process	High	Medium
Provision of evidence for systems that reuse existing components/subsystems	High	Low
Suitability and application of safety standards	High	Low

*Not applicable as the SLR did not collect these results

5. CONCLUSION

Managing safety evidence and presenting it clearly is an important but complex activity during the safety assurance and certification process. Despite extensive research on improving and supporting safety evidence management, only a small fraction of the past studies have been empirically validated in real project settings. Previous work further does not provide a broad picture of safety evidence provision and management activities, as the work has been limited to a small number of domains or countries. Subsequently, there is a lack of knowledge on what evidence items need to be constructed, how these items should be structured and assessed, and what challenges are associated with the process in practice.

This paper has presented the results of a questionnaire-based survey geared towards investigating the state of the practice on safety evidence management. The paper analysed 52 valid responses from 11

different domains and 15 countries, hence presenting an overall picture of the state of practice in safety evidence management. We analysed the industry perspective on the frequency of use of safety evidence types, evidence structuring and assessment techniques, and the challenges that practitioners face in evidence management. The paper analysed the commonalities among the reported domains in terms of evidence management, and compared its results against a previously performed state of the art study, discussing potential improvements for future research.

The results of our survey can be summarized as follows.

Respondents reported gathering evidence related to verification and validation artefacts such as *V&V Plan* and *Testing Results* very frequently. However some verification techniques such as *Model checking* and *Theorem proving* have been reported to be used in low numbers in the industry. Another interesting finding is that the evidence types concerning risk and hazard are not among the most frequently reported product-based types. A possible explanation could be that such information is normally embedded in *Requirements specification* (e.g., in the form of safety requirements or measures).

With regards to evidence change management, most respondents reported that they use manual techniques to check evidence completeness and to perform change impact analysis on evidence items. These results suggest a lack of tool support for completeness assessment and impact analysis.

With regards to safety evidence structuring techniques, non-graphical techniques such as *Textual Templates* and *Structured Text* seem to be used more 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.

As for safety evidence-assessment, the results suggest that *Checklists* and *expert judgment (with recorded rationale)* are the most common techniques. More studies on the reliability of expert judgment based safety assessment are an interesting area for future research. Although commonly reported for evidence assessment, the average rating of *Argumentation* for structuring evidence was lesser. Further studies to empirically compare the use of graphical notations and text-based argumentation for evidence structuring in practice might be interesting.

With respect to the challenges faced by the practitioners, the respondents shared a common sentiment, suggesting the existence of common and recurring set of obstacles faced in all domains.

A comparison of the results on the basis of application domain was performed for each research question. Briefly, we identified many commonalities in what evidence information is used, how it is structured, assessed and managed, and what obstacles are faced in the process. Some evidence types or techniques were used more in some domains than in others. All the domains acknowledged almost the same importance level for all the challenges provided. The result of the domain analysis could be used to compare the results of future studies on the state of the practice from different domains.

When comparing the state of the art and state of the practice, the results show that only a small set of evidence types (11 out of 49 items) have been given equal frequency in both literature and practice. Practitioners report evidence types that concern testing results to be more frequent, however, our results seem to suggest high contrast of difference when compared with the results from the systematic literature review. Differences were identified in the frequency of *Argumentation-based graphical notations*, with literature spending more effort on them and practitioners reporting them as less frequent. Similar contrast in the use of *Checklists* for evidence assessment was observed. The results suggest that a lot of research effort has been spent on techniques that have little adaptation in the industry. In general, the results indicate that researchers must try to promote more technology transfer from academia to industry and must try and evaluate the proposed approach in industrial settings.

The survey described in this paper is part of a larger research effort aimed at devising new tool support for safety evidence management activities. The insights gained from the survey are a stepping

stone for our future work activities, which include developing automation for traceability and impact analysis in safety evidence artefacts, and devising schemes for more systematic recording of expert judgment and using the rationale for more transparent evidence assessment.

ACKNOWLEDGMENT

The research leading to this paper has received funding from the FP7 programme under the grant agreement n° 289011 (OPENCROSS), 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 OPENCROSS partners who provided feedback on the questionnaire and that participated in instrument evaluation, and all the practitioners that participated in the survey.

Appendix A: Additional Domain Comparison Tables

This appendix presents the comparison among domains that were not part of the discussion in Section 4.6, due to low number of respondents. The comparison is presented in the form of tables in terms of evidence types (Table A - 1), techniques for evidence traceability (Table A - 2), structuring (Table A - 3) and assessment (Table A - 4) and challenges for evidence provision (Table A - 5).

Table A - 1. Percentage of respondents for each evidence type in the six application domains with low respondents

Evidence Types	Oil & Gas (3) %	Medical (2) %	Off-highway equipment (1) %	Nuclear (1) %	Mining (1) %	Machinery (1) %
Acceptance Testing Results	100	50	100	100	0	100
Accidents Specification	33	0	0	0	100	0
Activity Records	33	100	100	0	0	0
Architecture Specification	100	50	100	100	0	100
Assumptions and Conditions Specification	67	50	100	0	0	0
Automated Static Analysis Results	67	0	100	0	0	100
Communication Plan	0	0	100	100	100	0
Configuration Management Plan	33	50	100	0	0	100
Design Specification	100	50	100	100	0	100
Development and V&V Staff Competence Specification	33	0	0	0	0	100
Development Plan	100	50	100	0	0	0
Functional Testing Results	100	50	100	100	0	100
Hazards Causes Specification	67	50	100	0	100	0
Hazards Mitigation Specification	67	50	100	0	100	100
Hazards Specification	67	100	100	0	100	0
Inspection Results	33	50	100	100	0	0
Integration Testing Results	67	100	100	0	0	100
Model Checking Results	0	0	100	0	0	100
Modification Procedures Plan	33	50	100	0	0	100
Non-operational Testing Results	67	0	100	100	0	0
Normal Range Testing Results	67	100	100	100	100	100
Object Code	67	0	100	0	0	0
Operation Procedures Plan	67	50	100	0	100	100
Operational Testing Results	33	50	100	0	0	100
Operator Competence Specification	33	50	0	0	100	0
Performance Testing Results	100	50	100	100	100	100
Project Monitoring Plan	33	100	100	0	0	0
Reliability Testing Results	67	100	100	100	0	100
Requirements Specification	100	50	100	0	0	100
Reused Component Historical Service Data Specification	33	0	100	100	0	0
Reused Component Specification	67	50	100	0	0	100

Review Results	67	50	100	100	0	0
Risk Analysis Results	67	50	100	100	100	100
Risk Management Plan	33	50	100	100	100	0
Robustness Testing Results	67	0	100	0	0	100
Safety Management Plan	33	50	100	100	100	100
Simulation Results	67	0	100	100	0	100
Source Code	67	100	0	0	0	100
Stress Testing Results	33	0	100	0	0	100
Structural Coverage Testing Results	33	0	100	0	0	0
System Historical Service Data Specification	0	0	100	100	0	0
System Inception Specification	67	50	100	0	0	0
System Testing Results	33	100	100	100	0	100
Test Cases Specification	67	0	100	0	0	100
Theorem Proving Results	0	0	0	0	0	100
Tool Support Specification	33	0	100	0	0	100
Traceability Specification	33	50	100	100	0	100
Unit Testing Results	67	50	100	0	0	100
V&V Plan	67	50	100	100	0	100

Table A - 2. Percentage of respondents for each evidence traceability recording technique in the six application domains with low responses

Domain / (No. of respondents)	Evidence Traceability Techniques					
	Traceability Matrices	Models	Metadata	Hyperlinks	Naming conventions	Not recorded
Oil & Gas (3)	33	0	0	0	0	33
Medical (2)	0	0	0	0	0	50
Off-highway equipment (1)	0	100	0	100	100	0
Nuclear (1)	100	0	0	0	0	0
Mining (1)	0	0	0	0	0	0
Machinery (1)	0	100	0	100	0	0

Table A - 3. Average rating and mode of the frequency scale for each evidence structuring technique in the six application domains with low responses

Domain / (No. of respondents)	Evidence Structuring Techniques					
	Structured Text	Textual Template	Argumentation-Based Graphical Notation	Conceptual/Information Models	Process Models	Unstructured Text
Oil & Gas (3)	50% N/VO	33% N/VO/A	100% N	67% A	50% N/VO	50% N/VO
Medical (2)	50% R/S	50% R/VO	50% N/R	100% N	50% N/S	50% R/S
Off-highway equipment (1)	100% VO	100% VO	100% S	100% S	100% A	100% S
Nuclear (1)	100% S	0%	0%	0%	0%	100% VO
Mining (1)	100% VO	100% R	100% R	100% N	100% R	100% VO
Machinery (1)	100% R	100% R	100% VO	100% S	100% R	100% R

*A- Always; VO- Very Often; S- Sometimes; R- Rarely; N- Never

Table A - 4. Average rating and mode of the scale for each evidence assessment technique in the six application domains with low responses

Domain / (No. of respondents)	Evidence Assessment Techniques					
	Expert judgment without rationale	Expert judgment with rationale	Argumentation	Quantitatively	Qualitatively	Checklists
Oil & Gas (3)	50% S/A	67% S	67% VO	33% N/R/S	50% R/S	67% S
Medical (2)	50% S/VO	100% S	50% N/S	50% N/S	50% R/S	50% R/VO

Off-highway equipment (1)	100% N	100% S	100% Very Important	100% S	100% VO	100% A
Nuclear (1)	100% N	100% VO	100% S	100% R	100% VO	100% R
Mining (1)	100% S	100% VO	100% R	100% S	100% S	100% A
Machinery (1)	100% R	100% S	100% VO	100% A	100% S	100% S

*A- Always; VO- Very Often; S- Sometimes; R- Rarely; N- Never

Table A - 5. Average rating and mode of the scale for each evidence provision challenge faced in the six application domains with low responses

Domain / (No. of respondents)	Oil & Gas (3)	Medical (2)	Off-highway equipment (1)	Nuclear (1)	Mining (1)	Machinery (1)
<i>Compliance demonstration for new technologies</i>	33% OI/VI/I	100% MI	100% MI	0%	100% MI	100% VI
<i>Compliance demonstration for systems whose compliance has not been previously demonstrated</i>	33% UI/I/VI	50% MI/I	100% I	0%	100% I	100% VI
<i>Determination and decision upon the information that can be provided as evidence</i>	67% I	100% I	100% MI	0%	100% MI	0%
<i>Determination of confidence in evidence to support a particular claim about system safety</i>	67% VI	50% MI/I	100% VI	0%	100% I	100% I
<i>Existence of problems which, based on your experience, are exclusive to the application domain selected and do not arise in others</i>	67% MI	50% OI/I	100% I	0%	0%	100% I
<i>How to effectively create and structure safety cases</i>	33% MI/I/VI	100% I	100% I	0%	100% I	100% I
<i>Need for providing arguments to show how evidence meets the requirements/objectives of a safety standard</i>	67% VI	50% I/VI	100% VI	100% VI	100% MI	100% I
<i>Provision of adequate process information as evidence for the whole development and V&V process</i>	100% I	100% I	100% VI	100% VI	100% MI	100% I
<i>Provision of evidence for systems that reuse existing components/subsystems</i>	33% UI/I/VI	50% MI/I	100% I	100% I	100% MI	100% VI
<i>Suitability and application of safety standards</i>	67% VI	50% MI/I	100% I	0%	100% MI	100% MI

*VI- Very Important; I- Important; MI-Moderately Important; OI-Of little Importance; UI- Unimportant

Appendix B: Survey Questionnaire

EVIDENCE MANAGEMENT FOR COMPLIANCE OF CRITICAL COMPUTER-BASED SYSTEMS WITH SAFETY STANDARDS

Introduction

Most critical computer-based systems in domains such as avionics, railways, and automotive are subject to some form of safety assessment as a way to ensure that these systems do not pose undue risks to people, property, or the environment. The most common type of assessment is compliance with a safety standard. Examples of safety standards include IEC61508 for various types of systems, DO-178C for avionics, the CENELEC standards for railway, and ISO26262 for the automotive sector.

Demonstration of compliance with a specific standard involves gathering and providing convincing evidence of system safety. BY EVIDENCE, WE REFER TO THE INFORMATION THAT CONTRIBUTES TO DEVELOPING CONFIDENCE IN THE SAFE OPERATION OF A SYSTEM AND THAT IS USED TO MEET THE REQUIREMENTS/OBJECTIVES OF A SAFETY STANDARD. Examples of types of evidence are hazard analysis results, testing results, and reviews.

The aim of this survey is to gain insights into how practitioners manage evidence for demonstrating compliance of critical computer-based systems with safety standards. The survey has been designed as part of the work in OPENCROSS (<http://www.opencross-project.eu/>), a European research project on safety assurance and certification of critical systems. Among the aspects to research in OPENCROSS, the

Technical Report 2012-27, Simula Research Laboratory.

survey focuses on the information that is provided as evidence, how evidence change is managed, how evidence is structured, how its adequacy is assessed, and the challenges that can be faced to provide evidence.

The survey is targeted at PRACTITIONERS THAT DIRECTLY PARTICIPATE OR HAVE PARTICIPATED IN EVIDENCE MANAGEMENT FOR DEMONSTRATING COMPLIANCE OF CRITICAL COMPUTER-BASED SYSTEMS WITH SAFETY STANDARDS. The practitioners can correspond to people who have to provide evidence (e.g., an employee of a company that supplies components, such as a safety engineer or a tester), check others' evidence (e.g., an independent safety assessor), or request evidence (e.g., a person that represents a certification authority).

A questionnaire has been designed for completing the survey. Filling it is expected to take around 15 minutes. All the responses will be held confidential and anonymous.

Finally, if you are interested in the results of the survey, please contact Sunil Nair (sunil@simula.no) or Jose Luis de la Vara (jdelavara@simula.no).

Thank you very much for your participation in the survey.

Background Information

IMPORTANT: Background information must be completed in relation to your participation in the demonstration of compliance of critical computer-based system with safety standards.

1. What is the main application domain in which you are working regarding demonstration of compliance with safety standards? (IMPORTANT: ALL remaining questions must be answered in relation to the domain selected)

☐ Aerospace ☐ Automotive ☐ Avionics ☐ Defence ☐ Machinery ☐ Maritime ☐ Medical ☐ Nuclear ☐ Off-highway equipment
☐ Oil and gas ☐ Railways ☐ Robotics ☐ Telecommunications ☐ Trucks ☐ Other - please specify:

2. What are the safety standards for which you currently provide, check, or request evidence of compliance?

3. What country do you mainly work in regarding demonstration of compliance with safety standards?

☐ Australia ☐ Austria ☐ Belgium ☐ Brazil ☐ Canada ☐ China ☐ Finland ☐ France ☐ Germany ☐ India ☐ Italy ☐ Japan
☐ Netherlands ☐ Norway ☐ Poland ☐ Portugal ☐ Russia ☐ Spain ☐ Sweden ☐ UK ☐ USA ☐ Other - please specify:

4. What is the main role of the organization for which you work in the development of critical computer-based systems?

☐ Certification authority ☐ Component/system supplier ☐ Developer/manufacturer of final systems ☐ Independent safety assessor
☐ Regulation authority ☐ Development tool vendor ☐ Other - please specify:

5. How long have you been involved in activities related to demonstration of compliance with safety standards?

☐ Less than 1 year ☐ Between 1 and 2 years ☐ Between 2 and 5 years ☐ Between 5 and 10 years ☐ More than 10 years

6. How many projects targeted at demonstrating compliance with safety standards have you participated in?

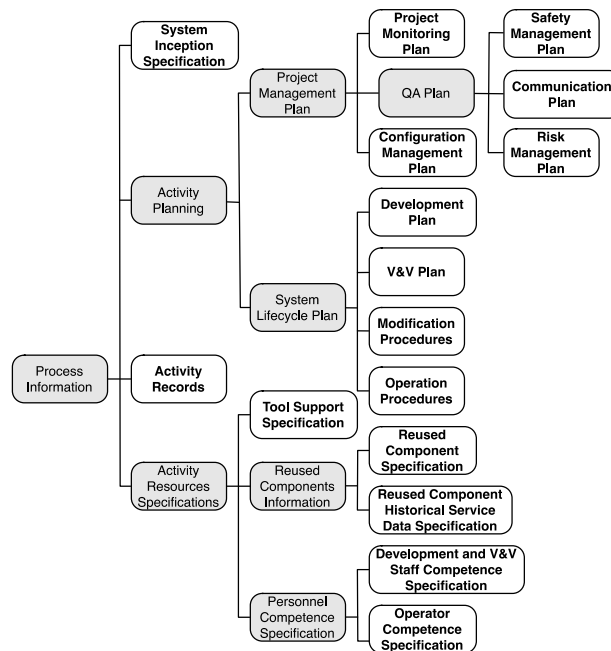
☐ Less than 5 projects ☐ Between 5 and 10 projects ☐ More than 10 projects

Page 3: Information Provided as Evidence (the page was randomized)

REMINDER: please answer the questions in relation to the application domain selected previously.

Safety evidence can be divided into process information (i.e., about the process followed to develop a critical system) and product information (i.e., about the characteristics of the system). Below, two figures show and classify different types of information (and artefacts) that might be used as process-based evidence and product-based evidence, respectively, for demonstrating compliance with safety standards. On this page you will be asked about the information provided, checked, or requested as evidence. More specifically, you will be asked about the leaf nodes of the classifications. Please note that SOME TYPES OF INFORMATION CAN BE REFERRED TO DIFFERENTLY in the application domain that you selected. You are kindly asked to read the definitions provided for each item carefully before deciding whether it applies to your domain or not.

PROCESS-BASED EVIDENCE

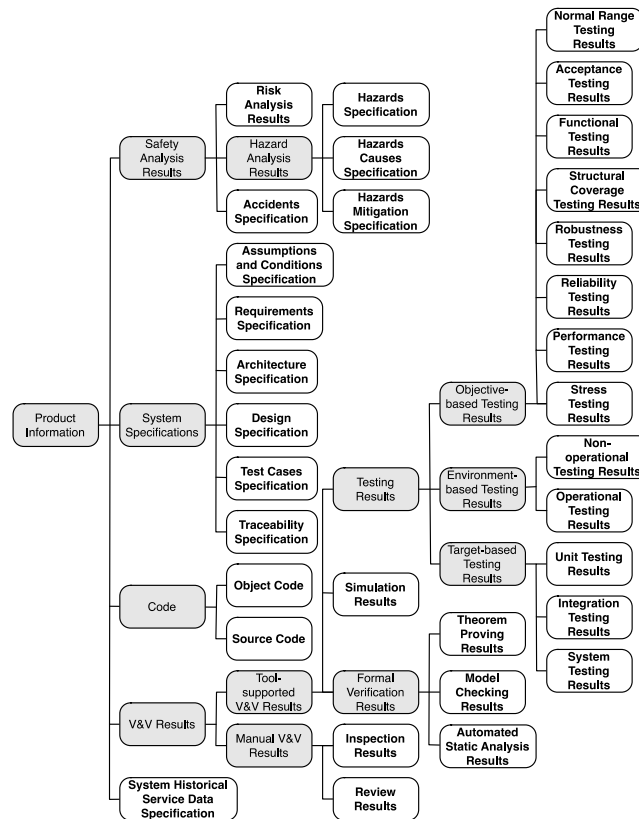


7. What process information (i.e., about the process followed to develop a critical system) do you provide, check, or request as evidence?

- ☐ **SYSTEM INCEPTION:** specification of initial details about the characteristics of a system and how it will be created.
- ☐ **PROJECT MONITORING PLAN:** description of how data about the actual progress of the activity planning of a system will be collected and compared with the baseline plans; e.g., meetings schedule and an organization chart.
- ☐ **SAFETY MANAGEMENT PLAN:** description of the coordinated, comprehensive set of processes designed to direct and control resources to optimally manage the safety of an operational aspect of an organization; e.g., safety culture and safety management processes.
- ☐ **COMMUNICATION PLAN:** description of the activities targeted at creating project-wide awareness and involvement in the development of a system; e.g., specification of the communication channels between service provider, device manufacturer, and regulation authorities.
- ☐ **PROJECT RISK MANAGEMENT PLAN:** description of the activity regarding the development and documentation of an organized and comprehensive strategy for identifying project risks; it includes establishing methods for mitigating risk and tracking them; e.g., risk reduction methodology.
- ☐ **CONFIGURATION MANAGEMENT PLAN:** description of how identification, change control, status accounting, audit, and interface of a system will be governed; e.g., version management and change control procedures.
- ☐ **DEVELOPMENT PLAN:** description of how a system will be built, which includes information about the requirements, design and implementation during coding and/or integration phases; e.g., development methodology and coding standards.
- ☐ **VERIFICATION AND VALIDATION PLAN:** description of how and by whom the verification and validation activities for a system will be executed; e.g., verification environment specification and tests plan.
- ☐ **MODIFICATION PROCEDURES PLAN:** description of the instructions about what to do when performing a modification in a system in order to make corrections, enhancements or adaptations to the validated system, ensuring that the required safety is sustained; e.g., change propagation and maintenance plan.
- ☐ **OPERATION PROCEDURES PLAN:** description of the instructions and manuals necessary to ensure that safety is maintained during system use; e.g., user manual and installation procedure.
- ☐ **ACTIVITY RECORDS:** artefacts collected during the execution of an activity planned for developing a system; e.g., maintenance log and review checklists.
- ☐ **TOOL SUPPORT SPECIFICATION:** specification of the different tools that will be used in the system lifecycle plan; e.g., tool qualification report.
- ☐ **REUSED COMPONENTS SPECIFICATION:** specification of the characteristics of an existing system that is (re)used to make up a system; e.g., reused component reliability specification and qualification documentation of a real-time operating system.
- ☐ **REUSED COMPONENTS HISTORICAL SERVICE DATA SPECIFICATION:** specification of the dependability of a component reused in a system based on past observation of the behaviour; e.g., mean time between failures.

- ☐ DEVELOPMENT AND V&V STAFF COMPETENCE: specification of the skills or knowledge that the parties involved in the development and V&V plans of a system need in order to perform the activities assigned to them; e.g., staff experience and tool training.
- ☐ OPERATOR COMPETENCE: specification of the skills or knowledge that the parties involved in the operation procedures need in order to perform the activities assigned to them; e.g., operational staff training needs specification.
- ☐ I do not provide, check, or request process information as evidence
- ☐ Other(s) – please specify:

PRODUCT-BASED EVIDENCE



8. What product information (i.e., about the characteristics of the system) do you provide, check, or request as evidence?

- ☐ RISK ANALYSIS/ASSESSMENT RESULTS: specification of the expected amount of danger when an identified hazard will be activated and thus become an accident in a system.
- ☐ HAZARDS SPECIFICATION: specification of the conditions in a system that can become a unique, potential accident.
- ☐ HAZARDS CAUSES SPECIFICATION: specification of the factors that create the hazards of a system.
- ☐ HAZARDS MITIGATION SPECIFICATION: specification of how to reduce hazard likelihood and hazard consequences when a hazard cannot be eliminated in a system.
- ☐ ACCIDENTS SPECIFICATION: specification of the conditions in a system that can become a unique, potential accident.
- ☐ ASSUMPTION AND CONDITIONS SPECIFICATION: description of the constraints on the working environment of a system for which it was designed.
- ☐ REQUIREMENTS SPECIFICATION: specification of the external conditions and capabilities that a system must meet and possess, respectively, in order to allow a user to solve a problem or achieve an objective, or to satisfy a contract, standard, or other formally imposed documents.
- ☐ ARCHITECTURE SPECIFICATION: description of the fundamental organization of a system, embodied in its components, their relationships to each other and to the environment, and the principles guiding its design and evolution.

Technical Report 2012-27, Simula Research Laboratory.

- ☐ DESIGN SPECIFICATION: specification of the components, interfaces, and other internal characteristics of a system or component.
- ☐ TEST CASE SPECIFICATION: specification of the tests inputs, execution conditions, and predicted results for a system to be tested.
- ☐ TEST RESULTS: results from the execution of test cases; they also indicate if the objectives and criteria of the tests have been met.
- ☐ TRACEABILITY SPECIFICATION: specification of the relationship between two or more pieces of information related to the development - process or product information - of a system.

☐ OBJECT CODE: computer instructions and data definitions in a form output by an assembler or compiler.

Source code: computer instructions and data definitions expressed in a form suitable for input to an assembler, compiler, or other translator.

☐ THEOREM PROVING RESULTS: results from the verification of a system by formally expressing its properties in a common language based on mathematical logic and using a theorem prover; a property can be shown to be a logical consequence of a set of axioms if it can be formally derived from the axioms with a set of deduction steps, which are instances of the set of inference rules that are allowed in the common language.

☐ MODEL CHECKING RESULTS: results from the verification of the conformance of a system to a given specification by providing a formal guarantee; the system under verification is modelled as a state transition system, and the specifications are expressed as temporal logic formulae that express constraints over the system dynamics.

☐ AUTOMATED STATIC ANALYSIS RESULTS: results from an automatic process for evaluating a critical system or component based on its form, structure, content, or documentation; e.g., static code analysis and cyclomatic complexity analysis.

☐ INSPECTIONS/AUDITS: results from the visual examination of system lifecycle products of a system in order to detect errors, violations of development standards, and other problems; e.g., code inspection.

☐ REVIEWS/WALKTHROUGHS: description of a process or meeting during which a work product or set of works products is presented to some interested party for comment or approval; e.g., design review.

☐ SIMULATION RESULTS: Results from the verification of a critical system by creating a model that behaves or operates like the system when provided with a set of controlled inputs; e.g., emulation and results from Matlab/Simulink.

☐ SYSTEM HISTORICAL SERVICE DATA SPECIFICATION: specification of the dependability of a system based on past observation of its behaviour; e.g., prior field reliability in similar applications.

☐ I do not provide, check, or request process information as evidence.

☐ Other(s) - please specify:

9. What types of testing are included in the product information (i.e., about the characteristics of the system) that you provide, check, or request as evidence?

☐ NORMAL RANGE TESTING: results from the verification of the behaviour of a system under normal operational conditions; e.g., equivalence classes and input partitioning testing.

☐ ACCEPTANCE TESTING: results from the validation of the behaviour of a system against the customers' requirements.

☐ FUNCTIONAL TESTING: results from the validation of whether or not the observed behaviour of a system conforms to its specification; e.g., hazard directed testing.

☐ STRUCTURAL COVERAGE TESTING: results from the verification of the behaviour of a system by executing all or a percentage of the statements or blocks of statements in a program, or specified combinations of them, according to some criteria; e.g., MC/DC and branch coverage testing.

☐ ROBUSTNESS TESTING: results from the verification of the behaviour of a system in the presence of faulty situations in its environment; e.g., fault injection testing.

☐ RELIABILITY TESTING: results from the verification of fault-free behaviour in a system; e.g., statistical and probabilistic testing.

☐ PERFORMANCE TESTING: results from the verification of the performance requirements of a system such as capacity and response time; e.g., timing and memory partitioning analysis.

☐ STRESS TESTING: results from the verification of the behaviour of a system at the maximum design load, as well as beyond it; e.g., boundary value and exhaustive input testing.

☐ NON-OPERATIONAL TESTING: results from evaluation of a system in an environment that does not correspond to but replicates its actual operational environment.

☐ OPERATIONAL TESTING: results from the evaluation of a system in its actual operating environment.

☐ UNIT/MODULE TESTING: results from the evaluation of the functioning in isolation of software pieces, which are separately testable; depending on the context, these could be the individual subprograms or a larger component made of tightly related units.

☐ INTEGRATION TESTING: results from the evaluation of the interaction between system components.

☐ SYSTEM TESTING: results from the evaluation of the behaviour of a whole system; external interfaces to other applications, utilities, devices, or the operating environment are also evaluated at this level.

Technical Report 2012-27, Simula Research Laboratory.

- ☐ I do not provide, check, or request testing information as evidence.
- ☐ Other(s) - please specify:

Page 4: Evidence Change Management (the page was randomized)

REMINDER: please answer the questions in relation to the application domain selected before.

A characteristic of evidence for demonstrating compliance with safety standards is that it can evolve. That is, a set of evidence can change because of, for instance, some modification in a system or the need to provide new evidence in order to guarantee system safety in a new context. This can affect single, isolated pieces of evidence as well as several pieces of evidence that are interrelated. For example, the modification of a requirement might affect the test cases specified to validate it. Consequently, the change of a piece of evidence can affect other pieces, which might become inadequate and/or might have to be (re)validated.

10. For the evidence that you provide, check, or request for demonstrating compliance with safety standards, how is the degree of completeness of evidence checked? **(the question and the options were randomized)**

- ☐ Manually (e.g., with a paper-based checklist)
- ☐ With tools that store and provide information about the degree of completeness for some types of evidence
- ☐ With tools that store and provide information about the degree of completeness for all types of evidence
- ☐ I do not know it

11. When a piece of evidence has changed, how is its effect on other pieces of evidence checked? **(the question and the options were randomized)**

- ☐ Manually, without following a predefined process
- ☐ Manually, according to a predefined process
- ☐ Automatically, using change analysis tools that provide information for the change effect of some types of evidence
- ☐ Automatically, using change analysis tools that provide information for the change effect of all types of evidence
- ☐ I do not know it
- ☐ Other(s) please specify:

12. Do you provide, check, or request details about how the change of a piece of evidence has affected others? **(the question and the options were randomized)**

- ☐ Yes ☐ No

13. In the documentation that you provide, check, or request for demonstrating compliance with safety standards, how is traceability between different pieces of evidence recorded? **(the question and the options were randomized)**

- ☐ Traceability matrices
- ☐ Models
- ☐ Metadata
- ☐ Hyperlinks
- ☐ Naming conventions
- ☐ Traceability between pieces of evidence is not recorded
- ☐ I do not know it
- ☐ Other(s) - please specify:

Page 5: Structuring of Evidence (the page was randomized)

REMINDER: please answer the questions in relation to the application domain selected previously.

14. This question lists a set of techniques that can be used for structuring evidence in order to show how it contributes to the fulfilment of the requirements/objectives of a safety standard. Please indicate how often you use, check, or request each technique (Never; Rarely; Sometimes; Very often; Always) **(the options were randomized)**

Technical Report 2012-27, Simula Research Laboratory.

- ☐ Unstructured text
- ☐ Structured text (providing patterns for the text to write)
- ☐ Textual templates (indicating the information to provide/the sections to fill)
- ☐ Argumentation-based graphical notations (e.g., GSN)
- ☐ Conceptual/information models (e.g., with UML)
- ☐ Process models (e.g., with SPEM)

15. If you would like to add any further techniques for structuring of evidence, please do so in the box below, and also indicate how often you use, check, or request them (for example, Technique X: very often; Technique Y: rarely, and so on)

Page 6: Evidence Adequacy Assessment (the page was randomized)

REMINDER: please answer the questions in relation to the application domain selected before

When managing evidence for demonstrating compliance with safety standards, it is also common to assess its adequacy. Adequacy is usually assessed based on the confidence in the information collected to support a particular claim about system safety. Adequacy can be estimated, for instance, by means of a qualitative approach (e.g., a level confidence) or a quantitative approach (e.g., a numerical estimation of the adequacy).

16. How often do you use, check, or request the following techniques for determining evidence adequacy? (Never; Rarely; Sometimes; Very often; Always) **(the options were randomized)**

- ☐ Expert judgement, without documenting the rationale behind the assessment
- ☐ Expert judgement, documenting the rationale behind the assessment
- ☐ Argumentation
- ☐ A quantitative approach (e.g., based on the use of Bayesian Belief Networks)
- ☐ A qualitative approach (e.g., based on the assignment of confidence levels to evidence)
- ☐ Checklists

17. If you would like to add any further techniques for evidence adequacy assessment, please do so in the box below, and also indicate how often you use, check, or request them (for example, Technique X: very often; Technique Y: rarely, and so on)

18. For the evidence that you provide, check, or request, do you check if the confidence in a piece of evidence is related to the confidence of other pieces? **(the question and the options were randomized)**

- ☐ Yes ☐ No

19. When a change occurs in the confidence in a piece of evidence that you provide, check, or request, do you check how the change might affect the confidence in other pieces of evidence? **(the question and the options were randomized)**

- ☐ Yes ☐ No

Page 7: Challenges in Evidence Provision (the page was randomized)

REMINDER: please answer the questions in relation to the application domain selected previously.

Practitioners might face different challenges when having to provide evidence for demonstrating compliance with safety standards. For example, safety standards can be difficult to understand, thus practitioners might have problems in determining what evidence has to be provided to comply with a safety standard

20. This question lists a set of possible challenges regarding provision of evidence for demonstrating compliance with safety standards. For those challenges that you have faced or observed, please indicate how important you consider them to be (Unimportant; Of little importance; Moderately important; Important; Very important) **(the options were randomized)**

- ☐ Compliance demonstration for new technologies (for example, model-driven technologies/development)
- ☐ Suitability and application of safety standards
- ☐ Determination and decision upon the information that can be provided as evidence

- ☐ Provision of adequate process information (i.e., about the process followed to develop a critical system) as evidence for the whole development and V&V process
- ☐ How to effectively create and structure safety cases
- ☐ Compliance demonstration for systems whose compliance has not been previously demonstrated (for example, a legacy system)
- ☐ Existence of problems which, based on your experience, are exclusive to the application domain selected and do not arise in others (for example, due to special regulations or processes)
- ☐ Determination of confidence in evidence to support a particular claim about system safety
- ☐ Need for providing arguments to show how evidence meets the requirements of a safety standard
- ☐ Provision of evidence for systems that reuse existing components/subsystems

21. If you would like to add any further challenges, please do so in the box below, and also indicate its importance (for example Challenge X: very important; Challenge Y: moderately important, and so on)

Follow-Up Studies

22. Finally, please fill the following information if you are interested in participating in follow-up studies (OPTIONAL)

- Name
- Organization
- Role
- Email

REFERENCES

- [1] Bozzano M, Villafiorita A. Design and safety assessment of critical systems: Auerbach Pub; 2010.
- [2] Kornecki A, Zalewski J. Certification of software for real-time safety-critical systems: state of the art. Innovations in Systems and Software Engineering. 2009;5:149-61.
- [3] IEC. Functional safety of electrical / electronic / programmable electronic safety-related systems (IEC 61508). 2005.
- [4] RTCA. DO-178C - Software Considerations in Airborne Systems and Equipment Certification. 2012.
- [5] Bate I, Kelly T. Architectural considerations in the certification of modular systems. Reliability Engineering & System Safety. 2003;81:303-24.
- [6] CENELEC. Railway applications - Safety related electronic systems for signalling, European Committee for Electrotechnical Standardisation CENELEC ENV 50129 2003.
- [7] ISO. International Standard Road vehicles — Functional safety - ISO/DIS 26262 2011.
- [8] Nair S, de la vara JL, Sabetzadeh M, Briand L. An Extended Systematic Literature Review on Classification, Structuring and Assessment of Evidence for Safety Compliance - Technical Report. 2013.
- [9] Wilson S, Kelly TP, McDermid JA. Safety case development: Current practice, future prospects. Safety and Reliability of Software Based Systems: Springer; 1997. p. 135-56.
- [10] Bouissou M, Martin F, Ourghanlian A. Assessment of a safety-critical system including software: a Bayesian belief network for evidence sources. Reliability and Maintainability Symposium, 1999 Proceedings Annual: IEEE; 1999. p. 142-50.
- [11] Bohner SA. Software change impact analysis. 1996.
- [12] Panesar-Walawege RK, Sabetzadeh M, Briand L, Coq T. Characterizing the chain of evidence for software safety cases: A conceptual model based on the IEC 61508 standard. Software Testing, Verification and Validation (ICST), 2010 Third International Conference on: IEEE; 2010. p. 335-44.

- [13] Baufreton P, Blanquart J, Boulanger J, Delseny H, Derrien J, Gassino J, et al. Multi-domain comparison of safety standards. Proceedings of the 5th International Conference on Embedded Real Time Software and Systems (ERTS2 2010), Toulouse, France (May 19-21, 2010)2010.
- [14] Ivarsson M, Gorschek T. A method for evaluating rigor and industrial relevance of technology evaluations. *Empirical Software Engineering*. 2011;16:365-95.
- [15] Kitchenham BA, Pfleeger SL. Personal opinion surveys. *Guide to Advanced Empirical Software Engineering*: Springer; 2008. p. 63-92.
- [16] Eriksson L-H. Using formal methods in a retrospective safety case. *Computer Safety, Reliability, and Security*: Springer; 2004. p. 31-44.
- [17] Torner F, Ohman P. Automotive Safety Case A Qualitative Case Study of Drivers, Usages, and Issues. *High Assurance Systems Engineering Symposium, 2008 HASE 2008 11th IEEE*: IEEE; 2008. p. 313-22.
- [18] Dodd I, Habli I. Safety certification of airborne software: An empirical study. *Reliability Engineering & System Safety*. 2012;98:7-23.
- [19] Liu S, Stavridou V, Dutertre B. The practice of formal methods in safety-critical systems. *Journal of Systems and Software*. 1995;28:77-87.
- [20] Panesar-Walawege RK, Sabetzadeh M, Briand L. Supporting the Verification of Compliance to Safety Standards via Model-Driven Engineering: Approach, Tool-Support and Empirical Validation. *Information and Software Technology*. 2012.
- [21] SafeCer. Deliverable D1.0.1 - State-of-practice and state-of-the-art agreed over workgroup. 2011.
- [22] OPENCROSS. D4.1 - Baseline for the common certification language 2012.
- [23] OPENCROSS. D5.1 - Baseline for the compositional certification approach 2012.
- [24] OPENCROSS. D7.1 - Baseline for the process-specific needs of the OPENCROSS platform. 2012.
- [25] OPENCROSS. D6.1 - Baseline for the evidence management needs. 2012.
- [26] Gerlach M, Hilbrich R, Weißleder S. Can Cars Fly? From Avionics to Automotive: Comparability of Domain Specific Safety Standards. *Proceedings of the Embedded World Conference*2011.
- [27] Robson C. *Real world research: A resource for social scientists and practitioner-researchers*: Blackwell Oxford; 2002.
- [28] Siegle D. *Likert Scale*. 2010.
- [29] Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A. *Experimentation in software engineering*: Springer; 2012.
- [30] Buckley FJ. *Implementing configuration management. Hardware, software, and firmware*. Los Alamitos, CA: IEEE Computer Society Press and Piscataway, NJ: IEEE Press; c1996, 2nd ed. 1996;1.
- [31] Garwood D. *Bills of Material: For a Lean Enterprise*: Dogwood Publishing, Incorporated; 2004.
- [32] Ericson CA. Introduction to System Safety. *Concise Encyclopedia of System Safety: Definition of Terms and Concepts*. 2011:1-15.

AUTHOR BIOGRAPHIES:

Sunil Nair Kolaserry Mohan is a PhD student at the the Certus Centre for Software V&V of Simula Research Laboratory since October 2011. He holds a Masters degree in Advance Software Eningeering from the University of Leicester (UK) and a Bachelors Degree in Computer Science and Engineering from Anna University (Chennai, India). His research interests include software engineering, safety assurance and certification, empirical software engineering, machine learning, automated testing and human judgement. He currntently serves as the program committee member in International Conference on Quality Software (QSIC) and SAFECOMP - Next Generation of System Assurance Approaches for Safety-Critical Systems (SASSUR).

Jose Luis de la Vara is a Postdoctoral Fellow at the Certus Centre for Software V&V of Simula Research Laboratory. He holds a PhD in Computer Science (2011) from the Universidad Politécnica de Valencia (Spain). His main research area is software engineering, with specific interest in requirements engineering, business process management, conceptual modeling, model-driven development, safety assurance and certification, and empirical software engineering. He has over 7 years of experience in research and collaboration with industry on these topics, and has published papers on them at top conferences such as CAiSE, ER, ICST, and REFSQ. He has been a member of the program committee or a reviewer of conferences and journals such as RE, BPM, SEKE, and IEEE Software.

Mehrdad Sabetzadeh received the PhD degree from the University of Toronto in 2008. He is currently a research scientist at the Interdisciplinary Centre for Security, Reliability and Trust (SnT) at the University of Luxembourg. From 2009 to 2012, he was a member of the research staff at Simula Research Laboratory, Norway. In 2009, he was a visiting researcher at University College London. His research interests are focused on model-based software engineering with a particular emaphsis on model-based verification and validation of business- and mission-critical applications.

Davide Falessi joined the Fraunhofer Center for Experimental Software Engineering in Maryland in 2012 as a Research Scientist in the Measurement and Knowledge Management Division. He currently serves as a program committee member in several international conferences including ESEM, WICSA, ICSR, SEKE, PROFES, EASE, and MTD. His main research interest is in devising and empirically assessing scalable solutions for the development of complex software-intensive systems with a particular emphasis on architecture, requirements, and quality. He received the PhD and the “Laurea” degrees in Computer Engineering from the University of Rome “TorVergata”.