Title: Validating FAROS risk models. Framework and application.

Deliverable n. 4.7

Task: 7.3
WP: 7
Responsible: AALTO – Jakub Montewka

Due delivery date: 2015-08-30
Actual delivery date: 2015-10-20

Dissemination level\(^1\): PU

\(^1\) PU = Public
PP = Restricted to other programme participants (including the Commission Services)
RE = Restricted to a group specified by the consortium (including the Commission Services)
CO = Confidential, only for members of the consortium (including the Commission Services)
Disclaimer

The information contained in this report is subject to change without notice and should not be construed as a commitment by any members of the FAROS Consortium. In the event of any software or algorithms being described in this report, the FAROS Consortium assumes no responsibility for the use or inability to use any of its software or algorithms. The information is provided without any warranty of any kind and the FAROS Consortium expressly disclaims all implied warranties, including but not limited to the implied warranties of merchantability and fitness for a particular use.

© COPYRIGHT 20012-2015 The FAROS Consortium.

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the FAROS Consortium. In addition, to such written permission to copy, acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.
Abstract

Validation of risk models is an open research question; which has not gained enough attention in the scientific literature yet. The risk models developed in the course of FAROS project are novel and tend to explore the areas not touched before. This implies, that the available knowledge on the issues addressed by the risk models is limited. Despite the existence of some empirical data on specific elements of the models, the data that would reflect the input and the output of the risk models is non-existent. In order to develop credibility in the risk models, mainly among the end-users, and to define the applicability boundaries of the models, we applied a validation framework that allows risk models evaluation in the presence of uncertainty and lack of data.

A validation framework adopted is based on three types of validity, which are prevalent in social sciences, see for example [1], [2]. Since the risk models developed in FAROS project revolve around human performance and specific factors affecting it, the adoption of validation framework from the social sciences seems reasonable. The following validity types are considered here: internal validity and construct validity, where the latter combines translation related validity and criterion related validity. To carry out the FCGRM validation process we extensively used the existing background knowledge obtained in the course of the literature survey performed in WP3 and new knowledge that has been gained in the course of WP 7. This new knowledge comes from the experiments using ship bridge simulators as reported in Deliverable 7.2.

As the results of validation tests, it is justified to conclude, that the risk models developed are plausible enough to serve as a basis for further reflections, allowing for the differentiation between various ship designs. Thus the risk models developed can be used at the initial stage of ship design process, as initially intended, allowing for relative comparison of various ship designs with respect to a predefine baseline design. In that mode, the risk models may be used by naval architects, vessel designers, and vessel system designers as intended, provided access to HF expertise is available to assist with application and interpretation. It is important to recognise the relevance of human factors input during its eventual application. HF provides the understanding of the complexities of human behaviour in operational settings, its interdependencies and interactions. However, it is not advised to apply risk models to the optimization process of ship design along with the utilitarian paradigm of decision making, involving cost-benefit analysis.
**Document Meta Data**

| Author/s: | Jakub Montewka, Doug Owen, Yasmine Hifi, Carlo Valbonesi, Romanas Puisa |
| In-house reviewers (optional): | Floris Goerlandt |
| Reviewer 1: |  |
| Reviewer 2: |  |
| Other reviewers: |  |
| Nature of Deliverable: | Report, Prototype, Demonstrator, Other |
| Related FAROS Deliverables: | D3.6, D4.6, D7.2 |

**Partners involved**

<table>
<thead>
<tr>
<th>No.</th>
<th>Organisation short name</th>
<th>Organisation full name</th>
<th>Name</th>
<th>email</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>AALTO</td>
<td>AALTO-KORKEAKOULUSAATIO</td>
<td>Jakub Montewka</td>
<td><a href="mailto:Jakub.montewka@aalto.fi">Jakub.montewka@aalto.fi</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Floris Goerlandt</td>
<td><a href="mailto:floris.goerlandt@aalto.fi">floris.goerlandt@aalto.fi</a></td>
</tr>
<tr>
<td>1</td>
<td>BBL</td>
<td>Brookes Bell R &amp; D</td>
<td>Romanas Puisa, Yasmine Hifi</td>
<td><a href="mailto:romanas.puisa@brookesbell.com">romanas.puisa@brookesbell.com</a>, <a href="mailto:yasmine.hifi@brookesbell.com">yasmine.hifi@brookesbell.com</a></td>
</tr>
<tr>
<td>3</td>
<td>DBL</td>
<td>Deep Blue SRL</td>
<td>Caro Valbonesi</td>
<td><a href="mailto:carlo.valbonesi@dblue.it">carlo.valbonesi@dblue.it</a></td>
</tr>
<tr>
<td>13</td>
<td>GAIN</td>
<td>Axencia Galega de Innovación CIS Galicia</td>
<td>Lucia Rodríguez Quiroga</td>
<td><a href="mailto:lucia.rodriguez.quiroga@xunta.es">lucia.rodriguez.quiroga@xunta.es</a></td>
</tr>
<tr>
<td>5</td>
<td>UoS</td>
<td>University of Strathclyde</td>
<td>Anthony Anderson</td>
<td><a href="mailto:tony.anderson@strath.ac.uk">tony.anderson@strath.ac.uk</a></td>
</tr>
<tr>
<td>12</td>
<td>LR</td>
<td>Lloyd's Register EMEA</td>
<td>Douglas Owen, Gemma Innes-Jones</td>
<td><a href="mailto:Doug.Owen@lr.org">Doug.Owen@lr.org</a>, <a href="mailto:Gemma.Innes-Jones@lr.org">Gemma.Innes-Jones@lr.org</a></td>
</tr>
</tbody>
</table>

**Document history**

<table>
<thead>
<tr>
<th>Version</th>
<th>Date of delivery</th>
<th>Changes</th>
<th>Author(s) Editor(s)</th>
<th>Reviewed by</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>15.08.2015</td>
<td>1st draft</td>
<td>J. Montewka</td>
<td>Y. Hifi</td>
</tr>
<tr>
<td>002</td>
<td>15.10.2015</td>
<td>2nd draft</td>
<td>D. Owen, Y. Hifi, C. Valbonesi</td>
<td>Y. Hifi</td>
</tr>
<tr>
<td>003</td>
<td>5.10.2015</td>
<td>3rd draft</td>
<td>J. Montewka</td>
<td></td>
</tr>
<tr>
<td>004</td>
<td>20.10.2015</td>
<td>Final version</td>
<td>R. Puisa, Y. Hifi</td>
<td>F. Goerlandt</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY ........................................................................................................... 6
PROBLEM DEFINITION ............................................................................................................ 6
TECHNICAL APPROACH ......................................................................................................... 6
RESULTS AND ACHIEVEMENTS ............................................................................................ 6
CONTRIBUTION TO FAROS OBJECTIVES ........................................................................ 6
1. INTRODUCTION AND OBJECTIVES .............................................................................. 8
   1 VALIDATION FRAMEWORK ............................................................................................. 9
      1.1 VALIDITY OF THE RISK MODEL ............................................................................ 11
         1.1.1 Nomological validity ......................................................................................... 11
         1.1.2 Internal validity ................................................................................................. 12
         1.1.3 Construct validity ............................................................................................. 12
   2 APPLICATION OF VALIDATION FRAMEWORK TO FAROS RISK MODELS ............ 14
      2.1 INTERNAL VALIDITY .............................................................................................. 14
         2.1.1 Internal validity related to FCGRM and FGPM ................................................. 14
         2.1.2 Internal validity related to FAROS fire risk model ........................................... 16
      2.2 CONSTRUCT VALIDITY RELATED TO COLLISION AND GROUNDING RISK
         MODELS ..................................................................................................................... 17
         2.2.1 Translation validity .......................................................................................... 17
         2.2.2 Criterion related validity ................................................................................... 22
      2.3 CONSTRUCT VALIDITY RELATED TO PERSONAL INJURY RISK MODELS .... 32
         2.3.1 Translation related validity .............................................................................. 32
         2.3.2 Criterion related validity - personal injury risk model ................................. 35
      2.4 CONSTRUCT RELATED VALIDITY - FIRE RISK MODEL .................................. 40
         2.4.1 Background ....................................................................................................... 40
         2.4.2 Translation related validity .............................................................................. 41
         2.4.3 Criterion related validity ................................................................................... 45
   3 SUMMARY AND CONCLUSION ...................................................................................... 46
      3.1 SUMMARY .............................................................................................................. 46
      3.2 CONTRIBUTION TO FAROS OBJECTIVES ........................................................... 47
      3.3 IMPLEMENTATION ................................................................................................. 47
   4 BIBLIOGRAPHY AND REFERENCES ............................................................................ 48
   5 INDEXES .......................................................................................................................... 50
      5.1 INDEX OF TABLES .................................................................................................... 50
      5.2 INDEX OF FIGURES ................................................................................................ 50
      5.3 ABBREVIATIONS ..................................................................................................... 50
Executive Summary

Problem Definition
Validation of risk models is a valid research question, however very often forgotten. Since the risk models tend to facilitate decision-making process, their validity is of high importance. In engineering validation is often tantamount with comparing the modelled outcome with the measurements. This approach is applicable when the measurements can be done, or when the data describing the modelled phenomena are available. However, in risk analysis, very often we face situation, that a risk model that is developed tends to describe a part of the world that has not been explored yet, thus neither data nor measurement is available to compare with. In such situation another approach to validate the model is required.

Technical approach
The risk models developed in the course of FAROS project are novel, centred on human performance under specific set of conditions, describing phenomena that have not been researched before. Thus, we are lacking the data that could be used for risk model validation. Due to that and human-centred models, we adopted a validation framework pertaining to social sciences, which evaluates risk models in a wider sense at several levels utilizing the available background knowledge, including data, theories, or experts judgments. Subsequently the models are evaluated at several levels, starting with the high level internal validity and then more detailed construct validity, which is further broken down into translation-related validity and criterion-related validity. Since the risk models are developed with the use of Bayesian Networks, also a set of specific tests is carried out pertaining to criterion/related validity. Finally the risk models are given a validation score, based on the adopted qualitative ranking. This shows in which areas the models are strong and where they need improvements.

Results and achievements
The objectives of this deliverable are as follows:

- To validate the risk models developed in WP4 and integrated in WP5, adopting the state-of-the art methods.
- To define the limits of the risk models application.
- To build-up the credibility in the risk models among the potential end-users.

The risk models presented here are developed to the best available knowledge, with the use of the state of the art modelling techniques and tools. The results of validation as reported in this document are good, since the models score in most cases moderate or good on three-level scale. The lower scores (poor or poor-moderate) are attributed to the type of validation tests, where a model is compared with some other models developed for similar purpose. Since the risk models developed here are unique, there exists no model that is developed for a similar purpose.

Contribution to FAROS objectives
Initially the purpose of the task described in this report was to cross-validate personal and societal (ship level) risk figures collected from the experimental data, with corresponding predications by risk models. However due to complexity of the topic analysed the quantitative findings from the experiment conducted remains inconclusive. Therefore the outcome of the risk models
cannot be compared with the experimental data. Moreover, due to adopted constructivist risk perspective, where the risk is seen in wider sense that just a number, the validation of risk models shall not be limited to benchmarking with the historical data or data obtained from empirical tests. This decision effectively decoupled the two sequential work-packages, WP6 (risk assessment and reduction) and WP7 (cross-validation), in terms of quantitative interaction and timing of completion\(^2\).

For that purpose, the wider validation framework was proposed and successfully applied to the risk models developed here. The framework allows model validation even in the lack of data, which was the case here. As demonstrated earlier the risk models developed are good representation of the objects of inquiry, which was proved by several validation tests. Therefore, the proposed validation framework clearly contributes to the project objectives in general and WP7 objectives in particular.

\(^2\) Technical arguments were provided to the EC who approved the change to the original plan.
1. Introduction and Objectives

The intention of this document is to validate the risk models developed in WP4 and WP5. The risk models developed in the course of FAROS project are novel and they try to measure quantities that have not been measured before. For this reason, conducting model validation in its classical way as usually done in engineering is not feasible, namely by comparing the models outcome with the historical trends or data. Therefore another form of validation needs to be adopted, where validity is concerned with the meaningfulness of risk model components. Therein the question of validity of a risk model is whether the risk model measures what it is intended to measure for its intended use, [1], [3].

Therefore, this deliverable introduces a validation framework assessing the plausibility of the risk model as a tool for serving its envisaged functions in the risk analysis. The model functions: (i) to convey an argumentation based on available evidence, (ii) to provide a basis for communication, (iii) to serve as an aid to thinking and (iv) to discriminate different ship designs. Risk models plausibility is evaluated using model-construct and risk-theoretical validity tests.

For this purpose a framework is adopted proposed in [4], which allows qualitative assessment of the validity of the FAROS Collision and Grounding Risk Model (FCGRM), FAROS Generic Personal Injury Model (FGPIM), and FAROS fire risk model, all developed with the use of Bayesian Networks. The validation framework is based on three types of validity, which are prevalent in social sciences, see for example [1], [5]. Since the risk models developed in FAROS project revolve around human performance and specific factors affecting it, the adoption of validation framework from the social sciences seems reasonable. The following validity types are considered here: nomological, internal and construct validity, where the latter combines translation related validity and criterion related validity.

To carry out the validation process we extensively used the existing background knowledge obtained in the course of the literature survey performed in WP3 and new knowledge that has been gained in the course of WP7. This new knowledge comes from the experiments using ship bridge simulators as reported in Deliverable 7.2. The objectives of this deliverable are as follows:

- To validate the risk models developed in WP4 and integrated in WP5, adopting the state-of-the art methods.
- To define the limits of the risk models application.
- To build-up the credibility in the risk models among the potential end-users.

This deliverable is structured as follows: Section 2 introduces a validation framework, which is applied to the FCGRM and FGPIM. The results of validation are provided and discussed in Section 3, whereas Section 4 concludes the deliverable.
1 Validation framework

Model validity is often conceptualized as a simple test of a model’s fit with a set of data. However validity is a much broader construct. In essence, validity is the ability of a model to describe the system that it is intended to describe both in the output and in the mechanism by which that output is generated, [4].

A major distinction between human performance models (HPMs) and physical system engineering models is the recognized need for HPMs to be sensitive to many complex aspects of application context as opposed to the representational austerity of physical models. This is true particularly because the primary human roles in advanced technology systems are to mediate the processing of information related to a broad range of context factors, [6]. This difference obviously complicates the model validation process.

Thus, model validation is much more than the simple comparison of model predictions with empirical data and the binary determination that the model is or is not valid. It is appropriate to view validation on a continuum of processes, [6].

At one end of the continuum is model calibration where we use the discrepancies between actual model predictions and empirical data to adjust parametric or structural aspects of the model in order to improve the correspondence for a subsequent execution of the same model. At the other end of the continuum, we have fundamental inquiry regarding the inherent value of different modeling frameworks, paradigms, and philosophies.

Between these two end-points lies the realm of practical model validations in various forms, depending on the degree of generality and scope of the model being validated.

As the risk models that are presented here are developed using Bayesian Belief Network (BBN), one needs to consider specificity of the model due to the modelling technique adopted. This specificity stems from the following elements that a BBN is built upon:

- model structure,
- node discretization,
- discrete state parameterization.

Therefore, the methodology that is applied here provides a framework to validate the risk models along the above-mentioned continuum; additionally the model specificity is recognized.

Validity concerns the question whether the analysis describes the specific concepts one intends to describe, for its intended use [7]. Therefore in this report we assess the plausibility of the risk model as a tool for serving its envisaged functions in the risk analysis. The risk models developed in the course of FAROS project functions to:

1. convey an argumentation based on available evidence,
2. provide a basis for communication,
3. serve as an aid to thinking, and
4. discriminate different ship designs.

Risk models plausibility is evaluated using model-construct and validity tests. The validation framework adopted here stems from the solid founded scientific discipline of
psychometrics, which attempts to measure the latent, unobserved variables, such as intelligence. Therein the validation framework is used to measure how well a particular test measure a latent variable. Psychometrics defines a several types of validity, which can be passed to varying degrees, providing a multidimensional measure of a test.

Since the risk qualifies for an unobserved variable, and the main actor of the risk models is human, the choice of the discipline of psychometrics to derive from is justified. The validation framework adopted here allows a multidimensional assessment of a risk model. Such an assessment allows definition of risk model limits and application areas, also the elements of the model that require further refinements can easily be pointed out.

The validation framework attempts to establish the nomological, internal and construct validity. High nomological validity indicates that the analysed model or test sits well within current academic thought on a subject. Internal validity assures that the causal mechanism adopted in the model is justified. Construct validity comprises of face and content validity, and describes to what extent a developed model describe a construct that is intended to operationalizes.

The validity types considered here are presented in Figure 1-1, and discussed further in the following sections. However, for more elaborate discussions the reader is referred to the cited publications.

![Figure 1-1: The structure of validity framework adopted here](image-url)
1.1 Validity of the risk model

As a reflection of a mind construct addressing possible consequences, which may or may not occur, a direct comparison between the risk model results and observations from the described system is not possible. In the accident reports, that have been analyzed the effect of Global Design Factors (GDFs) of Whole Body Vibration (WBV), noise, motions, and Deck Layout and Equipment arrangement and access (DLEAA) is not mentioned, [8], so a comparison with historic data is inconclusive. However, validity can be understood in a wider sense than a comparison with observed data, by inspecting the model qua model. Such approaches to validation are widely used in social science research [2], system dynamics modelling [9] and for expert-based Bayesian Network modelling [4].

First, it is possible to evaluate whether the model adequately operationalizes the construct it intends to measure, i.e. how well it concretizes the object of inquiry for the given purpose. This is evaluated in terms of face and content validity. Face validity (FV) is a subjective, heuristic interpretation of whether the model is an appropriate operationalization of the construct. Content validity (CV) is a more detailed comparison of the elements in the risk model in relation to what is believed to be relevant in the real system.

Second, a number of specific tests can be performed on the model, to evaluate whether the model adequately meets certain criteria. A behaviour sensitivity test (BST) is used to assess to which model elements the results are sensitive. The parameter sensitivity of a BN can be calculated and the results can be evaluated by domain experts. In a qualitative features test (QFT), the model response is evaluated for a number of test conditions in terms of a qualitative understanding of how the system is believed to respond under these conditions. In a concurrent validity test (CVT), the model elements are compared with the elements in another model for a similar purpose. This can also include a comparison with the output of such a model if the scope of the applications is the same.

The validity tests do not “prove” that the model results are correct, but only indicate the extent to which the model is a plausible representation of the object of inquiry. This relates to adopted understanding of risk and the adopted risk perspective, where no reference is made to an underlying “true” risk. The model should be plausible enough to serve as a basis for further reflections, leading to deliberative judgments.

1.1.1 Nomological validity

Nomological validity provides the evidence that the structural relationships among variables/constructs is consistent with other studies that have been measured with validated instruments and tested against a variety of persons, settings, times, and, methods.

---

3 More model validity tests have been proposed in the literature then the ones retained here, e.g. a dimensional consistency tests, boundary adequacy tests and structure verification tests [9]. Which tests are considered largely depends on the type of developed model. For the purposes of this report, a limited number of relatively straightforward tests is retained.
The nomological map, which is drawn based on nomological distance between available and relevant models, may serve as a reference for choosing appropriate comparison models, see [4].

1.1.2 Internal validity
Internal validity is the approximate truth about inferences regarding cause-effect or causal relationships. This type of validity is relevant in studies that try to establish a causal relationship. The key question in internal validity is whether observed changes can be attributed to intervention (i.e., the cause) and not to other possible causes (sometimes described as "alternative explanations" for the outcome). It reflects the extent to which a causal conclusion based on a study is warranted. Since the causality is a fundamental premise of the risk models developed in FAROS, therefore this type of validity test is of high importance here.

1.1.3 Construct validity
Once the internal validity is established, and the structure of the model is developed, its validity can be judged. For this purpose construct validity is applied, which refers to how well the analyst translates a concept or an idea – that is a construct – into a functioning and operating reality that is the operationalization. Construct validity comprises four types of validity: face, content, concurrent and predictive, as depicted in Figure 2-1 and described further in the following sections. These types are grouped in two categories: translation validity and criterion-related validity, which are discussed in the following subsections, for elaborated discussion on these the reader is referred to [5].

1.1.3.1 Translation validity: face and content validity
Translation validity focuses on whether the operationalization reflects the true meaning of the construct. It attempts to assess the degree to which constructs are accurately translated into the operationalization, using subjective judgment – face validity – and examining content domain – content validity, [1].

Face validity is a subjective judgment on the operationalization of a construct. In other words, this type of validity checks, whether or not the model that is validated captures the phenomenon it is intended to capture. It is commonly accepted, that this is a crude and basic measure of validity, [10].

For the models that are developed with the use of Bayesian Belief Networks (BBN), to evaluate the face validity, the following set of questions can be asked [4]:
- Does the model structure (the number of nodes, node labels and arcs between them) look the same as the experts and/or literature predict?
- Is each node of the network discretized into sets that reflect expert knowledge?
- Are the parameters of each node similar to what the experts would expect?

Content validity can be defined as a qualitative type of validity where the domain of the concept is made clear and the analyst judges whether the measures fully represent the domain, [11].

To test for content validity of a model one can check that all the relevant factors and relationships from the literature are included in it. Moreover, all the relationships, which are novel to the model, need to be discovered.
To assess the content validity of a BBN model, the following questions can be asked [4]:

- Does the model structure contain all and only the factors and relationships relevant to the model output?
- Does each node of the network contain all and only the relevant states the node can possibly adopt?
- Are the discrete states of the nodes dimensionally consistent?
- Do the parameters of the input nodes and CPT reflect all the known possibilities from expert knowledge and domain literature?

### 1.1.3.2 Criterion related validity

Criterion-related validity is the degree of correspondence between a test measure and one or more external, so called "objective", criteria. The "objective" results are obtained either by a well-established instrument ("the gold standard") or by direct measurement. The criterion validity may be quantified by the correlation coefficient between the two sets of measurements.

Depending on the nature of the reference data set, the criterion validity measures are subdivided into:

- concurrent validity and predictive validity
- convergent and discriminant validity measures.

#### Concurrent and predictive validity

An example of concurrent validity test is a comparison of the results obtained from FCGRM with the results from the experiments. Alternatively the model elements are compared with the elements in another model for a similar purpose.

FCGRMs measures influence of GDFs on the risk of collision, grounding and injury through the quantification of the performance of a navigator in a given task. By comparing the effect of GDFs on the performance of a navigator with the results obtained from a ship simulator, one can evaluate the level of concurrent validity of the relevant parts of FCGRMs.

Concurrent validity in the context of BBN can refer to the possibility that a network or section of a network behaves identically to a section of another network, preferably driven by data. To test the concurrent validity of the structure of a BN, we can check other networks in related domains for sub-networks that are similar to sub-networks in our model.

The following questions are suggested as tests of a BNN's concurrent validity, [4]:

- Does the model structure or sub-networks act identically to a network or sub network modelling a theoretically related construct?
- In identical sub networks, are the included factors discretized in the same way as the comparison model?
- Do the parameters of the input nodes and CPTs in networks of interest match the parameters of the sub network in the comparison model?

Predictive validity refers to the ability of a model to estimate some event or outcome in the future. A number of specific tests can be performed on a model, to evaluate whether the model adequately meets certain criteria. A behavior sensitivity test (BST) is used to assess to which model elements the results are sensitive. The parameter sensitivity of a BN can be calculated and the results can be evaluated by domain experts. In a qualitative features test (QFT), the model response is evaluated for a number of test conditions in terms of a qualitative understanding of how the system is believed to respond under these conditions.
Predictive validity in a BBN based model can be considered to encompass both the model behaviour and the model output. The following can help to establish confidence in the predictive validity of BBN: behaviour sensitivity tests, qualitative features analysis and the extreme conditions tests, [4].

- The behaviour sensitivity test can be applied to the model structure and parameters by determining to which factors and relationships the model is sensitive, and comparing this to hypothetical models or alternative empirical models.
- Qualitative features analysis is a case of predictive validity testing where behaviour in a hypothetical model is compared to the behaviour of individual pairs of nodes, subnetworks and the entire model, [12].
- The extreme conditions test can be seen as a special case of qualitative features analysis, as it sets the hypothetical model to extreme conditions where the behaviour of the model is more predictable.

To perform the predictive validity test, the following questions can be addressed, [4]:

- Is the model behaviour predictive of the behaviour of the system being modelled?
- Once simulations have been run, are the output states of individual nodes predictive of aspects in the comparison models?
- Is the model sensitive to any particular findings or parameters to which the system would also be sensitive?
- Are there qualitative features of the model behaviour that can be observed in the system being modelled?
- Does the model including its component relationships predict extreme model behaviour under extreme conditions?

In answering the questions related to validity tests, one can assess the adequacy of the models using a qualitative scale

- Solid supporting evidence that a model performs well: GOOD
- Some supporting evidence that a model performs well: MODERATE
- Little or no supporting evidence that a model performs well: POOR

As with the BN construction itself, the answers to the validation tests need to be justified.

### 2 Application of validation framework to FAROS risk models

This section demonstrates the use of validation tests introduced earlier on FAROS risk models and the results obtained. The tests tend to increase the credibility of the risk models developed, by providing thorough explanation of risk models and their behaviour. Moreover they demonstrate that the risk models can be successfully used in the process of risk-based ship design, which is their primary aim.

#### 2.1 Internal validity

##### 2.1.1 Internal validity related to FCGRM and FGPIIM

One of the aims of collision and grounding risk models was to describe the process through which exposure to GDFs causally affects the probability of the specified unwanted outcomes aligned with the mechanism derived from attention management
theory and incident data. To achieve this, a causal pathway was required that linked the input (exposure to GDFs) with the output (unwanted outcomes) through the mediating agent of the crewmember. For the elaborate description of the risk models, the reader is referred to FAROS deliverable 4.6, see [13]

Importantly, the causal chain represents the effects of GDF exposure on human performance in a way that could be developed and elaborated in the risk model. It served to do three things:

1. Represent the mechanism by which GDF exposure impacts collision and grounding risk
2. Describe the overall topography of the final model
3. Facilitate the identification of nodes

The casual path from the exposure to GDFs, though human behaviour to the occurrence of an unwanted outcome of a collision or grounding, is summarised in Figure 2-1.

Figure 2-1 Causal chain describing the relationship between crew GDF exposure and unwanted outcomes

Two main paths linking GDF exposure to human behaviour (and subsequently to collision and grounding) have been identified:

- **Path 1: Stressor effects.** Exposure to a GDF acts as a stressor and can affect the perceptual, cognitive and physical capabilities of an individual (e.g. attention management), which can subsequently impair the performance of the individual (i.e. the actual behaviour produced).

- **Path 2: Physical effects.** Exposure to a GDF can have specific and direct effects on the behaviour produced. For example, Ship motion can result in Motion-Induced Interruptions (MII). MII does not affect the underlying human capabilities of balance or fine motor control, but it exceeds the ability of the
human to compensate and produce the intended behaviour. Similarly, WBV can directly impact the actual behaviour produced.

These two paths show how GDF exposure affects human behaviour, which in turn influences the performance of safety critical tasks. It is the outcomes of and individual’s actions and behaviour that determine the success or failure of a safety critical task. Insufficient performance of the safety critical tasks associated with maintaining safe vessel navigation and avoiding collision or grounding create an antecedent for the unwanted outcome. However, insufficient task performance alone does not determine whether or not a collision or grounding occurs; the vessel must also be exposed to the collision or grounding hazard. For example:
- For a collision to occur, another vessel must also be present and abstain from taking collision avoidance action.
- For a grounding to occur, the ship must be in shallow water.

Without exposure to these hazards, a collision or grounding will not occur. It is the combination of insufficient task performance with exposure to these hazards that may result in the unwanted outcomes of the collision or grounding of the vessel. As such, the FAROS risk model represents a plausible causal mechanism of GDF exposure leading to unwanted outcomes.

**2.1.2 Internal validity related to FAROS fire risk model**

In this section we address the fire ignition model for machinery spaces, as explained in Table 8. The structure of the model can be represented as shown below.

![Figure 2-2 Structure of the ignition model (only main directly influencing factors are shown)](image)

The red box represents probabilities of flammable oil and high temperature surfaces which were modelled by combining the elicited empirical data from marine engineers and a Bayesian network to quantity the human error. The assumed causal chain that involves the human error is shown below and one can see that the human error led to increased susceptibility to fire due to prolonged availability of heat and fuel sources.
There were two human error types considered:

- The first one is related to failure to properly detect and assess a problem. This probability is generic, i.e. space and ship independent, and denoted as $P(\text{HE.DA})$, with the acronym DA standing for detection and assessment.
- The second type is related to any human error made during maintenance of ship systems. It is also called maintenance performance failure (MPF) probability, which is denoted as $P(\text{HE.MPF})$.

These probabilities are linked to the global design factors (GDFs) considered in the project, and the links are implemented by means of a Bayesian network shown in Figure 2-14. Note, $P(\text{HE.DA})$ and $P(\text{HE.MPF})$ probabilities are calculated in the two corresponding yellow nodes. Specifically,

$$P(\text{HE.DA}) = P(\text{Insufficient detection & assessment task performance}), \quad (1)$$

whereas

$$P(\text{HE.MPF}) = P(\text{Insufficient maintenance task performance}) \quad (2)$$

### 2.2 Construct validity related to collision and grounding risk models

#### 2.2.1 Translation validity

The scientific literature was examined in the pursuit of quantitative data that describes the relationship between exposure to GDFs and human performance. Some quantitative data was found to inform the relationship between some of the components of the causal chain, this was not possible for all of the links, [13], [14].

It was found that the data on the specific GDF effects of ship motion (with the exception of MII), noise, WBV and DLEAA on human performance are sparse and in many (but not all) cases generated under very specific, often non-marine, conditions. What data there are shows that there is certainly evidence for GDFs having some effect on human performance.
performance. However, the direct effects of GDF exposure on human performance tend to be weak, whereas secondary effects acting through another mechanism (e.g. fatigue, MIS) tend to be stronger and more pervasive.

Specifically there are some data that describe the:
- Impact of GDFs on specific human capabilities
- Impact of GDFs on specific human behaviours
- Impact of errors on task performance
- Relationship between hazard exposure and collision/grounding

However, there is very little data about the link between the following components:
- Degraded human capabilities and collision or grounding related performance
- Degraded task performance and exposure to the collision / grounding hazard

Figure 2-4 highlights the links in the causal chain for which some quantitative data are available (in green) and the links for which there is no data (in red).

**Figure 2-4 Supporting data for links in the causal chain, [13].**

Despite the lack of quantitative data to describe the causal mechanism, one of the major findings in WP3 was the identification of a theoretical framework to explain the impact of GDFs on human behaviour (see FAROS deliverable D3.5 for a comprehensive description of this theoretical framework).

The approach that emerged from WP3 combined the principles from three theoretical models:
1. Dynamic Adaptability Model, [15],
2. Cognitive Control Model, [16]
3. Malleable Attentional Resources Theory, [17].

Taken together these theories describe a mechanism that accounts for the impact of stressors on human performance, based on the principles of attention management.

Under the DAM paradigm, GDFs are seen as types of physical stressor that affect human capabilities associated with maintaining a desired level of task performance either directly or indirectly (e.g. via fatigue). When exposed to GDFs, CCM describes humans compensating through the effortful direction of more cognitive resources at the task, typically at the cost of performance in other areas. Despite the sophisticated (and potentially subconscious) strategies humans have at their disposal, there is a limit to how much an individual can compensate without experiencing degradation in primary or secondary task performance.

In addition, the extent to which human can compensate for task demands is not fixed. MART describes this compensatory capability changing as a function of task demands and associated arousal an individual experiences – attentional resources available vary as a function of load. When humans are in a state of under-load (i.e. bored) their pool of attentional resources is relatively small and will increase proportionately with the demands placed on them. However, there is a limit to how much the pool of attentional resources can grow. When task demands exceed the pool of attentional resources available (either transiently or when the upper attentional resource limit is exceeded), performance can breakdown and errors may be made.

Generally, task performance is only expected to degrade and become insufficient when compensatory mechanisms have failed. However, the literature does not allow prediction of how and when (chronologically) an operator would fail, under what conditions of GDF exposure, and what the specific effect on behaviour (i.e. type of error) would be.

In terms of risk modelling, an approach based on attention management theory allows representation of the effect of GDF exposure as a stressor that sits either above or below the threshold of attentional capacity for any given task. If the stressor exceeds the attentional capacity then a negative effect is expected, whereas no negative effect on human performance would result if the stressor can be managed within the available attentional capacity.

Representing ship motion, noise and WBV GDFs as stressors interacting with an individual's attention management capabilities provides an evidence-based mechanism for human performance that has been used to develop the FAROS risk models.

The approach was guided by the relevant theoretical models available in the scientific literature. However, the ship bridge simulator studies performed within WP4 and WP7 provided some useful information, which supports the approach taken.

The HRA method Nuclear Action Reliability Assessment (NARA) was selected to provide the HEPS associated within collision and grounding model. NARA is a third generation HRA method that, while nuclear industry focussed, uses a broad range of
industries in the CORE-DATA dataset underlying the HEP calculation, arguably making it more suitable for navigation tasks performed on the bridge. NARA was adopted to enhance the accuracy of the risk model through the generation of validated (albeit non-marine specific) HEPs associated with task characteristics that are compatible with tasks performed by the Officer of the Watch (OOW) and helmsman. NARA also provided baseline error rates for a given Generic Task Type (GTT) unaffected by GDFs. This allowed probabilistic estimation of the effect of GDF exposure on HEPs.

NARA categorises the factors that negatively influence human performance as one of eighteen Error Producing Conditions (EPCs). The EPC that best represented the causal mechanism from GDF exposure to human performance was EPC No. 15: Poor Environment. This EPC represents the stressor effect of GDF exposure on attention management capability. The potential strength of effect of this EPC was set using the Assessed Proportion of Affect (APOA) variable. The APOA level was set based on the application of the NARA methodology to subjectively determine an appropriate value, nominally between 0 (no effect) and 1 (maximum effect). However, based on the guidance available for NARA, it was decided to cap the maximum APOA associated with the EPC to 0.1.

2.2.1.1 Face validity

The following section summarises the face validity of the FAROS collision and grounding risk models with respect to the questions defined by Pitchforth & Mengersen (2013).

Q1: Does the model structure (the number of nodes, node labels and arcs between them) look the same as the experts and/or literature predict?

The face validity of the models is drawn from the evidence base that drove the development of the model. The major components that make up the model of the attention management mechanism and calibration against available accident data were derived from the literature, as documented in FAROS deliverables D3.3 and D3.4.

Q2: Is each node of the network discretized into sets that reflect expert knowledge?

The input node thresholds for effect on human performance were derived from the literature, however the probabilities were set by a combination of expert judgement and human reliability assessment (HRA) methods. The HRA HEART method selected is evidence-based to the extent it utilises the CORE-DATA set to set basic human error probabilities that are derived from real-world human reliability data. However, significant expert judgement is required to select which factors are appropriate for the context and to determine the magnitude of their effect on human performance in the context. Despite the adoption of HRA methods in a number of domains (e.g. nuclear, oil & gas), these subjective judgements required by the method remain subject to potential inter-rater variability.

Q3: Are the parameters of each node similar to what the experts would expect?

There are very few models inhabiting the space of the models presented here, hence it is difficult to elaborate on their consistency with the parameters of FAROS models. Some elements of FAROS models correspond to the expert driven risk models for collision and grounding. Structurally, the concept of degraded and normal attention and resulted human behaviour are the aspects the FAROS models have in common with a
model proposed by DNV, (Hänninen & Kujala 2012). However, the relationship of the attention management construct to other elements in the model is unique to the FCGRM. Overall, the face validity of the FCGRM is high.

### 2.2.1.2 Content validity

The following section summarises the content validity of the FCGRM with respect to the questions defined by Pitchforth & Mengersen (2013).

**Q1: Does the model structure contain all and only the factors and relationships relevant to the model output?**

The development of the FCGRM was based on scientific literature as far as possible. The pre-selected GDFs of ship motion, noise, whole body vibration (WBV) are a subset of the potential performing-shaping factors that could influence the outcomes of collision or grounding. As no prior screening process was performed to select the GDFs included in the FAROS project, other factors can be imagined that are influenced by vessel design which have a greater impact on human performance.

**Q2: Does each node of the network contain all and only the relevant states the node can possibly adopt?**

The states for the GDF input variables of ship motion, noise and vibration were derived from the literature around thresholds of effect that, despite incomplete data, can be justified with some caveats. Other states for nodes within the model represent a simplification due incomplete knowledge (attention management being represented by only two states: normal or degraded), or to perform a function within the model (switching GDF stressor effects and GDF physical effects on or off). These binary nodes are understood to have low content validity.

**Q3: Are the discrete states of the nodes dimensionally consistent?**

According to the definition by Pitchforth & Mengersen [4], nodes outside the GDF input variables fail the test of dimensional consistency as they do reflect states either above or below a level, but cannot be set at the level itself. Hence these reduce the content validity of the model.

**Q4: Do the parameters of the input nodes and conditional probability tables (CPTs) reflect all the known possibilities from expert knowledge and domain literature?**

The FCGRM has been developed to the limits of the relevant literature and expert knowledge. The integration of HRA is relevant here. Despite HRAs general limitations as a method in terms of reliability and its application here being outside its normal use, it was necessary as a means to generate probabilistic bounds for nodes within the FCGRM as described in FAROS D4.6.

Overall, the content validity of the FCGRM is moderate, being compromised by the limitations imposed by the limited scientific knowledge in the domain.
2.2.2 **Criterion related validity**

2.2.2.1 **Concurrent validity**

Concurrent validity (CV) of risk models developed in FAROS project is tested in two-fold. First, by comparing the elements of FCGRM with another model developed for a similar purpose. Second by comparing the results obtained from FCGRM with the results from the experiments.

The former is complicated, since there is no other model that would link the GDFs with the collision and grounding probability. The external risk models that were used for this validation exercises, were developed for specific purposes, other than the purpose of FAROS. Therefore the inter-models comparison is feasible only with respect to selected elements of models. For instance the elements describing human behaviour in a given situation can be compared with respect to the paradigms governing these, discretization level adopted or input-output relation. In some case qualitative comparison is feasible; in other quantitative evaluation is the only option.

Below there is the most recent list of relevant collisions and grounding risk models, with a brief characteristics of each. Subsequently the first step of concurrent validity is conducted.

**Comparison with other, relevant risk models**

A model by Leva et al. estimates the performance of a bridge operator, where the contextual factors are accounted for. The Authors adopted a theoretical framework estimating the probability of navigator error while in an encounter with other ship, called CREAM, [18].

A model by DNV estimates the probability of collision, accounting for the navigational parameters, safety culture, personnel factors, management factors, technical reliability and other vigilance, [19]. The model assumes the failure of a navigator along the path of detection-assessment-action for an accident to happen. As a modeling technique BBN is adopted, and experts judgments is used as a primary source of information to parameterize the model. The model was developed to measure the effects of bridge layout on the collision and grounding probability through the estimation of performance of a navigator. The performance affects the causal path of: detection, assessment and action, moreover, the probability of technical failure of ship machinery is accounted for, likewise in FCGRM. Therefore, the nomological distance between this model and FCGRM is relatively short.

A model by Martins and Maturana (2010) quantifies the human failure contribution to the collision and grounding of oil tankers, where the effect of possible combinations of performance shaping factors is investigated. A set of fault trees are used to develop the model, which is parameterized with the use of THERP, [20]. In a later work of the same authors, a methodology based on BBN for analyzing human reliability is presented and applied to the operation of an oil tanker, focusing on the risk of collision accidents. The model adopts BBN as a modeling tool, unlike the earlier work the model is not based on any theoretical foundations when it comes to the quantification of human error probability. Since the aim of the model is not the actual computed risk value but rather the determination of elements that considerably impact the system risk, [21]. In nomological sense these models and FCGRM seem to be distal.
A model by Asami and Kaneko estimates the probability of collision avoidance failure in such a way that the effect of some risk control options can be evaluated quantitatively (installation of AIS, radar redundancy, presence of coastal pilot, VHF radio phone). It uses the naturalistic decision making framework and the concept of situation awareness, see [22]. Quantification of the failure probability is made with the use of the Technique for Human Error Rate Prediction (THERP). The nomological distance between this model and FCGRM is large.

For the CV (concurrent validity) test, we compare FCGRM with the risk models, which we found compatible and which are closest in the nomological sense. However only one risk model by DNV is found relevant for this purpose. Since FCGRM and DNV model are developed with the use of BBN, the CV test is performed, answering the following questions, pertaining to the models developed with the use of BBN, see [4]:

- **Q1:** Does the model structure or sub-networks act identically to a network or sub network modelling a theoretically related construct?
- **Q2:** In identical sub networks, are the included factors discretized in the same way as the comparison model?
- **Q3:** Do the parameters of the input nodes and CPTs in networks of interest match the parameters of the sub network in the comparison model?

**Q1: Does the FCGRM structure or its sub-networks act identically to the DNV model’s network or sub network?**
Since FCGRM is very specific and DNV model is wide, these two are not fully comparable. DNV model encompasses crew competence, bridge layout, target detection is split into visual and through technical means, and also the effect of third parties on the assessment and action is estimated. However, if the DNV model is stripped from those, and the core is retained, where the effect of stressors on the performance of a navigator and resulting probability of collision or grounding is estimated, these two models can be compared at a generic level.
Since the basic causal relations: stressor – attention of a person – detection – assessment – action is retained in both models, the logic behind their structure can be seen as similar.
Therefore the answer to the question is positive.

**Q2: In identical sub networks, are the included factors discretized in the same way as the comparison model?**
Since the two models are developed with two different paradigms behind, in regards to human reliability, and the types of stressors and their effects, some of the parameters in the analysed sub networks are discretized differently.
In DNV model the process of detection-assessment-action is divided into three variables, two of them are binary and one (assessment) has three states (correct, wrong, no assessment). FCGRM models this process through one, binary node.
Similarly the node called Attention in DNV’s model has three states (high, low, not able to pay attention), whereas FCGRM’s node called Attention Management Capability has two states (normal, degraded).
Even though the nodes compared are not discretised in the same manner (three states versus two states), the logic behind the discretization is similar and the two utmost states of the three-states variable correspond to two-states variable. Moreover, both models calculate the probability of an accident, which is governed by the utmost state of the compared variables.
Therefore, the answer to Q2 is positive.

**Q3: Do the parameters of the input nodes and CPTs in FCGRM match the parameters of the sub network in the comparison model?**

Due to different modelling paradigms, there is only one parameter that matches in both models, that is *Steering failure*. However, the CPTs behind this node differ in the two risk models.

The structure of FCGRM is based on the concept of Attention Management and FCGRM’s parameters are found through the HRA methods and experts judgments. The DNV model to large extent is based on experts’ judgment and historical data, therefore the structure and parameters of these two models can hardly be comparable. Despite the negative answer to this question, the validity of the FCGRM is not compromised. Since the purposes of two compared models are substantially different, their input nodes and CPTs are not expected to match.

**Comparison with simulator studies**

FCGRM measures influence of GDFs on the risk of collision, grounding and injury through the quantification of the performance of a navigator in a given task. By comparing the effect of GDFs on the performance of a navigator with the results obtained from experiments with the use of ship bridge simulator, one can evaluate the level of concurrent validity of the relevant parts of FCGRM.

The results of simulator studies conducted in the course of FAROS are reported in Deliverable 7.2, [23]. However the studies did not account for the effect of vibration nor motion on human performance, which due to technical limitations of the simulator could not be realistically modelled. In the report authors claim that there are no significant effects of noise or task difficulty on either of the navigation measures (normalised distance to the relevant target ship, and normalised track deviation). However, when the instructors’ ratings of the rapidity/quality of mariners’ responses are considered, a number of significant or marginally significant effects of noise and task difficulty emerge. The overall, global ratings of task performance by instructors showed both a significant effect of task difficulty (which validated the task difficulty manipulation within the experiment) and a significant effect of noise such that higher levels of noise were associated with poorer performance. The effects of noise were also either significant, or close to statistical significance, in the case of several specific actions within scenario A (one of the two ‘high difficulty’ scenarios), suggesting that there might be an interaction between noise and task difficulty such that noise has little or no effect in situations where the task is easier, but might lead to impaired performance of a high difficulty task. Despite the lack of quantitative data, the qualitative evaluation of FCGRM and the subjective and qualitative results of simulator studies are feasible here. These show that there exists an effect of noise on human performance, especially in case of difficult tasks. These findings support, to some extent, the modelling choices adopted while developing FAROS collision and grounding risk models.

The overall concurrent validity of the FCGRM can be judged as between poor and moderate.
Figure 2-5 DNV collision model
Figure 2-6 FAROS collision risk model
2.2.2.2 Predictive validity

Predictive validity refers to the ability of a model to estimate some event or outcome in the future. A number of specific tests can be performed on a model, to evaluate whether the model adequately meets certain criteria. A behaviour sensitivity test (BST) is used to assess to which model elements the results are sensitive. The parameter sensitivity of a BN can be calculated and domain experts can evaluate the results. In a qualitative features test (QFT), the model response is evaluated for a number of test conditions in terms of a qualitative understanding how the system is believed to respond under these conditions. Finally, the results of the models from the nomological proximity can be compared.

The behaviour sensitivity test

BST can be applied to the model structure and parameters by determining to which factors and relationships the model is sensitive. This type of test, together with the uncertainty assessment, provides a valuable tool for screening the model for important variables, which are both uncertain and the model is sensitive to. By ranking the variables in the model by their importance, we define the set of variables affecting the most the credibility of the risk model.

Based on the findings from the sensitivity analysis, the following can be concluded, [25]:

- Collision and grounding risk models are highly sensitive to the following parameters: Maintenance Task Performance, C1 - Detection, Assessment and execution of simple actions and D1 - verbal communication of safety critical data.
- The model assessing the probability of collisions is also sensitive to Evasive action of another ship and Helmsman present. However, the effect that these parameters have on the output is significantly lower than the effects of C1 and D1, as specified above.
- The model assessing the probability of grounding is sensitive to Helmsman present.
- The remaining nodes have very low sensitivity values, meaning that their effects on the models outputs are rather minor.

Secondly, the evidential uncertainty assessment has been carried out on the most sensitive model parameters. To rank the uncertainty, the following qualitative scoring system is applied, see [26]:

**Significant uncertainty**

One or more of the following conditions are met:

- The phenomena involved are not well understood; models are non-existent or known/believed to give poor predictions.
- The assumptions made represent strong simplifications.
- Data are not available, or are unreliable.
- There is lack of agreement/consensus among experts.

**Moderate uncertainty**

Conditions between those characterising significant and minor uncertainty, e.g.:

- The phenomena involved are well understood, but the models used are considered simple/crude.
- Some reliable data are available.
Minor uncertainty
All of the following conditions are met:

- The phenomena involved are well understood; the models used are known to give predictions with the required accuracy.
- The assumptions made are seen as very reasonable.
- Much reliable data are available.
- There is broad agreement among experts.

Table 1: The qualitative assessment of evidential uncertainty for models assessing the probability of an accident, [25]

<table>
<thead>
<tr>
<th>Model parameter</th>
<th>Justification for the evidential uncertainty score</th>
<th>Evidential uncertainty score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance Task Performance</td>
<td>This node represents the performance of navigation tasks critical in collision or grounding avoidance and provides a structure compatible with the introduction of a NARA GTT, potentially affected by EPC No. 15 via ‘Attention Management Capability’ and GDF Physical effects</td>
<td>Moderate</td>
</tr>
<tr>
<td>C1 - Detection, Assessment and execution of simple actions</td>
<td>This node represents the performance of vessel manoeuvring instructions critical in collision or grounding avoidance when the helmsman present and introduction of a NARA GTT, potentially affected by EPC No. 15 via ‘Attention Management Capability’.</td>
<td>Moderate</td>
</tr>
<tr>
<td>D1 - verbal communication of safety critical data</td>
<td>The node represents the communication of vessel manoeuvring instructions critical in collision or grounding avoidance when a helmsman present and introduction of a NARA GTT, potentially affected by EPC No. 15 via ‘Attention Management Capability’.</td>
<td>Moderate</td>
</tr>
<tr>
<td>Evasive action of another ship</td>
<td>The node represents the performance of navigation tasks critical in accident avoidance on board other ship. It is quantified based on NARA.</td>
<td>Moderate</td>
</tr>
<tr>
<td>Helmsman present</td>
<td>At the moment this node is quantified fully based on judgement. However, more detailed assessment is possible, by performing survey among shipping companies.</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

NARA was also used to determine a baseline human error rate (Nominal HEP) to set the task performance HEP unaffected by GDF exposure. One of the limitations of the application of NARA in this context was highlighted by the selection of the task from the predefined list within NARA from which the HEP was established to represent task performance associated with the hazard exposure. To limit the complexity of the model, a single GTT was sought to represent all relevant navigational tasks performed by the OOW that are important in managing collision or grounding risk. The GTT that is most analogous is:

Task C1 – Simple response to alarms/indications providing clear indication of situation (Simple diagnosis required) Response might be direct execution of simple actions or initiating other actions separately assessed. (Nominal HEP = 0.0005)

A second GTT was identified to account for possibility that a helmsman may also be present. In this case the helmsman is steering the ship based on verbal instructions communicated by the OOW. The GTT that is most analogous is:

Task D1 - Verbal communication of safety critical data

While having a helmsman present may introduce the possibility of a miscommunication error with the OOW, NARA also recognises a mitigating effect of a team. The NARA
Human Performance Limiting Value for ‘Actions taken by a team of operators’ was used to cap the potential error rate at $1\times10^{-4}$ for the condition where a helmsman is present. The same value is taken for the probability of potential error of not performing evasive action by another ship involved in the encounter.

The NARA calculation allows inclusion of multiple EPCs and an Extended Time Factor (ETF). In this risk model for collision and grounding, GDFs are represented using only one EPC and there is little justification to include the ETF. Thus, the HEP was calculated based on the following formula:

$$\text{HEP} = \text{GTT} \times [(\text{EPC} - 1) \times (\text{APOA} + 1)]$$

By combining the results of sensitive and uncertainty assessment, the parameter importance ranking is carried out. For the models presented here, three parameters have high importance score, namely Maintenance Task Performance, C1-Detection, Assessment and Execution of Simple Action, D1 – verbal communication on safety critical data.

Table 2: The qualitative assessment of model parameters importance for models assessing the probability of an accident [25].

<table>
<thead>
<tr>
<th>Model parameter</th>
<th>Evidential uncertainty score</th>
<th>Sensitivity score</th>
<th>Importance score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance Task Performance</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>C1 - Detection, Assessment and execution of simple actions</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>D1 - verbal communication of safety critical data</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Evasive action of another ship</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Helmsman present</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Since there are three elements which are both uncertain and the FCGRM is sensitive to, we decided to modify the model to accommodate the available alternative hypothesis that can describe the analyzed variables. Separate nodes, called alternative hypothesis (AH), were added to the model, each reflecting the potential values for a given variable, with associated probabilities. The AH will not affect the level of sensitivity of a variable that it addresses, but it attempts to address the epistemic uncertainty associated with a given variable. By adopting the AHs a point estimate for the probability of collision / grounding is obtained, given the a’priori distribution of the values that each AH takes. Alternatively a distribution of the probabilities of collision / grounding is obtained, by executing the model n-times, each time sampling the values of AH from a predefined distribution.

**Qualitative features analysis**

QFA is a case of predictive validity testing where behaviour in a hypothetical model is compared to the behaviour of individual pairs of nodes, subnetworks and the entire model, [12]. A special case of QFA is the extreme conditions test, which sets the hypothetical model to extreme conditions where the behaviour of the model is more predictable. The results of these tests are presented in Table 3, where the probabilities of collision and grounding are shown for three different states of input variables. The base line refers to the input variables being in their states reflecting the average
conditions on board ships. The all-active state refers to a situation where all the stressors are active and are affecting negatively the performance of a navigator. The all-inactive state refers to the situation where there is no stressor acting on a navigator, thus his performance is not affected. The same results, however in non-dimensional form, are depicted in Figure 2-7, where the relative changes with respect to the base line are calculated.

Table 3 The probabilities of collision and grounding obtained in the course of extreme conditions tests.

<table>
<thead>
<tr>
<th></th>
<th>The probability of collision</th>
<th>The probability of grounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>All active</td>
<td>1.38E-06</td>
<td>3.58E-04</td>
</tr>
<tr>
<td>Base line</td>
<td>1.26E-06</td>
<td>3.06E-04</td>
</tr>
<tr>
<td>All inactive</td>
<td>7.01E-07</td>
<td>2.11E-04</td>
</tr>
</tbody>
</table>

Figure 2-7 The results of extreme conditions tests performed on collision and grounding risk models.

Comparison of the results obtained from the models being in nomological proximity

There is only one risk model that is nomologically close to the FCGRM, namely DNV risk model. In this section we present the results of comparison test between FCGRM and DNV model. The proximity of these two models comes from the fact, that both measure the effect of performance shaping factors (PSFs) on the ability of a navigator to perform his main tasks in relation to safe navigation and collision avoidance action. However, the PSFs that each model takes into account are different, likewise the mechanisms governing that ability. In FCGRM navigator’s ability is governed by the attention management capability, which is modelled by a variable of the same name. In DNV model this ability is modelled through a set of variables, where the one called performance seems to be the most comparable with the attention management in FCGRM. Therefore the effect of manipulation of these two variables on the explanatory variable (the probability of collision) is measured. The results of this test are presented
in Figure 2-8, in non-dimensional form, referring the base-line level. In Table 4 the probabilities of collision associated with the various states of the analysed variables are shown.

It is evident, that the behaviour of both models is comparable, when changing the states of the variables of interests between their states. In FCGRM the variable *attention management capabilities* has two states (Normal, Degraded), however in DNV model the variable *performance* has four states (Excellent, Standard, Poor, Not able to perform). For comparison we took *performance=Excellent* and *performance=Poor*, these refer to normal and degraded attention management capability (AMC).

![Figure 2-8 The non-dimensional results of comparison of two nomologically close collision risk models: FCGRM and DNV model.](image)

**Table 4 The results of comparison of two nomologically proximate collision risk models.**

<table>
<thead>
<tr>
<th>States of selected nodes in the risk models</th>
<th>DNV</th>
<th>FCGRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collision</td>
<td>Collision</td>
<td>Collision</td>
</tr>
<tr>
<td>Increase performance</td>
<td>8,1E-06</td>
<td>1,0E-06</td>
</tr>
<tr>
<td>performance=excellent / AMC= normal</td>
<td>8,60E-06</td>
<td>1,26E-06</td>
</tr>
<tr>
<td>Deteriorate performance</td>
<td>9,5E-06</td>
<td>1,4E-06</td>
</tr>
<tr>
<td>performance=poor / AMC=degraded</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The overall predictive validity of the FCGRM can be judged as between moderate and high.
2.3 Construct validity related to personal injury risk models

2.3.1 Translation related validity

The structure of the FAROS Generic Personal Injury Model (FGPIM), shown in Figure 2-9 below, was based on the FHPM described in section 2.1 above. The associated BBN is shown below in Figure 2-10, where an additional performance shaping factor (PSF), DLEAA (deck layout equipment arrangement and access) is considered. The effect of DLEAA, is assessed in terms of its potential to affect the human performance by constituting a constraint on the possible ways of doing the work. In addition the notion that the location of areas may affect also time pressure was encapsulated within the risk model in the assessment criteria for the DLEAA ‘Colocation of Areas’ factor.

Figure 2-9 Structure of the FAROS Generic Personal Injury Model (FGPIM)
2.3.1.1 Face validity

The following section summarises the face validity of the FGPIM with respect to the questions defined by Pitchforth & Mengerse, [4].

Q1: Does the model structure (the number of nodes, node labels and arcs between them) look the same as the experts and/or literature predict?

The face validity of the FGPIM is drawn from the evidence base that drove the development of the model. The major components that make up the model of the attention management mechanism and calibration against available accident data were derived from the literature, as documented in FAROS deliverables D3.3 and D3.4.

Q2: Is each node of the network discretized into sets that reflect expert knowledge?

The input node thresholds for effect on human performance were derived from the literature, however the probabilities were set by a combination of expert judgement and human reliability assessment (HRA) methods. The HRA HEART method selected is
evidence-based to the extent it utilises the CORE-DATA set to set basic human error probabilities that are derived from real-world human reliability data. However, significant expert judgement is required to select which factors are appropriate for the context and to determine the magnitude of their effect on human performance in the context. Despite the adoption of HRA methods in a number of domains (e.g. nuclear, oil & gas), these subjective judgements required by the method remain subject to potential inter-rater variability.

Q3: Are the parameters of each node similar to what the experts would expect? There are very few models inhabiting the space of the FGPIM, hence if it difficult to establish to what extent the FGPIM reflects consistency with node parameters defined by other experts. Structurally, however, the concept of safe and unsafe behaviours is an aspect the FGPIM has in common with a model proposed by Tomas et al. (1999). However, the relationship of the safety behaviour construct to other elements in the model is unique to the FGPIM.

Overall, the face validity of the FGPIM is high.

2.3.1.2 Content validity

The following section summarises the content validity of the FGPIM with respect to the questions defined by Pitchforth & Mengersen, [4].

Q1: Does the model structure contain all and only the factors and relationships relevant to the model output? The development of the FGPIM was based on scientific literature as far as possible. However, the demands that BBNs have for probabilistic data based on discrete thresholds or continuous data plotting an effect placed limits on the type of factors that could be incorporated into the model. In this case, for example, fatigue effects from exposure to GDFs could not be incorporated as literature provides neither threshold values for its onset nor a curve describing the effect of GDF exposure on fatigue.

The pre-selected GDFs of ship motion, noise, whole body vibration (WBV), and deck layout equipment arrangement and accessibility (DLEAA) are a subset of the potential performing-shaping factors that could influence the outcomes of collision, grounding, personal injury and fire. As no prior screening process was performed to select the GDFs included in the FAROS project, other factors can be imagined that are influenced by vessel design which have a greater impact on human performance.

Q2: Does each node of the network contain all and only the relevant states the node can possibly adopt? The states for the GDF input variables of ship motion, noise and vibration were derived from the literature around thresholds of effect that, despite incomplete data, can be justified with some caveats. This is also the case for the incident type and personal injury outcome nodes.

Other states for nodes within the model represent a simplification to due incomplete knowledge (attention management being represented by only two states: normal or degraded), or to perform a function within the model (switching GDF stressor effects and
GDF physical effects on or off). These binary nodes are understood to have low content validity.

**Q3: Are the discrete states of the nodes dimensionally consistent?**

According to the definition by Pitchforth & Mengersen [4], nodes in the FGPM outside the GDF input variables fail the test of dimensional consistency as they do reflect states either above or below a level, but cannot be set at the level itself. Hence these reduce the content validity of the model.

**Q4: Do the parameters of the input nodes and conditional probability tables (CPTs) reflect all the known possibilities from expert knowledge and domain literature?**

The FGPM has been developed to the limits of the relevant literature and expert knowledge. The integration of HRA is relevant here. Despite HRAs general limitations as a method in terms of reliability and its application here being outside its normal use, it was necessary as a means to generate probabilistic bounds for nodes within the FGPM as described in FAROS D4.5.

Overall, the content validity of the FGPM is moderate, being compromised by the limitations imposed by the limited scientific knowledge in the domain.

### 2.3.2 Criterion related validity - personal injury risk model

#### 2.3.2.1 Concurrent validity

The concurrent validity as introduced earlier refers to the possibility that certain subnetworks of the actual system are similar to other subnetworks representing similar systems.

The questions associated with determining concurrent validity as defined in [4] are:

- **Q1**: Does the model structure or sub-networks act identically to a network or sub network modelling a theoretically related construct?
- **Q2**: In identical sub networks, are the included factors discretized in the same way as the comparison model?
- **Q3**: Do the parameters of the input nodes and CPTs in networks of interest match the parameters of the sub network in the comparison model?

The challenge here is to identify candidate subnetworks to compare the personal injury risk model to. By the nature of the problem that the personal injury risk model addresses in FAROS it is almost impossible to define the nomological map of the risk model.

During the development of the personal injury risk model, the literature review revealed a very limited number of models relevant to the project. The only model using Bayesian Networks modelling technique addresses the problem of injuries due to fall from heights in the building industry. This model has only the modelling technique in common with the personal injury risk model of FAROS and is not a good candidate for assessing the concurrent validity of the model.

Another model developed for occupational accidents in the offshore industry was the closest in term of field of study but used a different modelling technique that would not make the comparison meaningful.
Finally a model looking at accidents occurring in a workplace using Structural Equation Modelling (SEM) could be a potential candidate for comparison by [27] referred to hereon in as Tomas’s model.

In terms of modelling technique, the SEM is very close to BNs[^4], but in this case the comparison can only be done at a high level. The discussion will be limited to the structure of the network as the discretisation and the parametrisation cannot be compared. As such, only question 1 from [4] above will be addressed. Tomas’s hypothetical model is presented in Figure 2-11 below

**Figure 2-11 The hypothetical Tomas’s model**

Tomas's model and the FGPIM both have a “main linear” path leading to unwanted outcomes. In Tomas model they are “accidents”, and in the FGPIM they are a “personal injury”.

When looking at both models, we could see some similarities in the subnetwork formed by the variables: Safety Behaviour, Actual Risk, Accidents and Hazards for Tomas’s model and the subnetwork formed by the Safety behaviour, Incident types, Hazard exposure and Personal injury for the FGPIM.

---

[^4]: http://www.mii.ucla.edu/causality/?p=571
In the context of Tomas's model, the *Safety Behaviour* was measured on a 12-item scale. This scale asked workers about the existence and frequency of safe and unsafe behaviours on the job (correct use of machines, observance of safety rules, speed at work, alcohol ingestion, etc.).

The *Actual Risk* is seen as the workers’ perception of the real probability of suffering an accident. It is the level of hazards that still remains after taking into account all safe and unsafe behaviours.

However, the occurrence of *Accidents* as a dependent variable on *Safety Behaviour* and *Actual Risk* was found to be problematic by the authors. They used instead a composite measure of occurrence of accidents. This was defined as the sum of the number of accidents and the score for the severity of the last three accidents.

*Hazards* was defined as a variable that measures the participant’s perceived level of intrinsic hazards at the work site. It was assumed that the Hazard variable was an exogenous variable (a variable that is independent from the state of the other variables in the causal model).

In the FGPIM, *Safety behaviour* refers to the behaviour required by the individual to keep them safe whilst performing a task. While the specific safety behaviour required is dependent upon the task (e.g. holding the handrail (safety behaviour) when going down stairs (task)), all tasks will have behaviours and actions that are required to keep an individual safe.

*Incident types* lists the most common incident types associated with personal injury and death on passenger and tanker vessels.

*(Injurious) Hazard exposure* refers to whether or not the individual has been exposed to a sufficient amount of hazard (e.g. kinetic, thermal, electrical, etc.) to result in an injury. It is the combination of ‘insufficient’ safety behaviour, exposure to an incident type, and exposure to a hazard at a level that in injurious that determines whether or not an individual is injured. This node captures the fact that not all incidents result in injury.

*Personal injury* is the result of exposure to an injurious hazard. Distinction is made only between three levels of personal injury outcome: no injury, injury and fatality.

The above subnetworks can only be compared at a high level due to the differences in the elements of the main linear causal path. As the underlying principals behind both subnetworks, seem to be quite similar, the answer to question 1 is positive.

The major difference though is that all variables of the FGPIM subnetwork are dependent on other variables, whereas in Tomas’s model, the variable representing Hazards is independent from the rest, increasing the nomological distance.

Overall, the concurrent validity of the FGPIM is low to moderate because of limited models available for comparison.
2.3.2.2 Predictive validity

As for the collision and grounding model, a number of specific tests were performed on a model to evaluate whether the model adequately meets certain criteria. These are summarised below.

The behaviour sensitivity test

Table 5 summarises the findings from the sensitivity analysis documented in FAROS deliverable D4.5.

Table 5. Node groupings by sensitivity

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly sensitive</td>
<td>Safety Behaviour</td>
</tr>
<tr>
<td></td>
<td>Injurious Hazard Exposure</td>
</tr>
<tr>
<td></td>
<td>Incident Type</td>
</tr>
<tr>
<td>Sensitive</td>
<td>Attention Management</td>
</tr>
<tr>
<td>Somewhat sensitive</td>
<td>Room Geometry [Note: Represents ‘Space Allocation’, ‘Location of Areas’ and ‘Accessibility / Circulation’ nodes]</td>
</tr>
<tr>
<td>Less sensitive</td>
<td>Motion - Lateral Acceleration</td>
</tr>
<tr>
<td>Insensitive</td>
<td>All other nodes</td>
</tr>
</tbody>
</table>

Following the sensitivity analysis, an evidential uncertainty analysis was performed. The evidential uncertainty analysis qualitatively assessed the evidence underpinning each node to which outcome of the FGPM is most sensitive. The assessment was based on the criteria described in Table 6.

Table 6. Uncertainty assessment rating criteria

<table>
<thead>
<tr>
<th>Uncertainty rating</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant</td>
<td>One or more of the following conditions are met:</td>
</tr>
<tr>
<td></td>
<td>• The phenomena involved are not well understood; models are non-existent or known/believed to give poor predictions.</td>
</tr>
<tr>
<td></td>
<td>• The assumptions made represent strong simplifications.</td>
</tr>
<tr>
<td></td>
<td>• Data are not available, or are unreliable.</td>
</tr>
<tr>
<td></td>
<td>• There is lack of agreement/consensus among experts.</td>
</tr>
<tr>
<td>Moderate</td>
<td>Conditions between those characterising significant and minor uncertainty, e.g.:</td>
</tr>
<tr>
<td></td>
<td>• The phenomena involved are well understood, but the models used are considered simple/crude.</td>
</tr>
<tr>
<td></td>
<td>• Some reliable data are available.</td>
</tr>
<tr>
<td>Minor</td>
<td>All of the following conditions are met:</td>
</tr>
<tr>
<td></td>
<td>• The phenomena involved are well understood; the models used are known to give predictions with the required accuracy.</td>
</tr>
<tr>
<td></td>
<td>• The assumptions made are seen as very reasonable.</td>
</tr>
<tr>
<td></td>
<td>• Much reliable data are available.</td>
</tr>
<tr>
<td></td>
<td>• There is broad agreement among experts.</td>
</tr>
</tbody>
</table>

The evidential uncertainty analysis was performed on the nodes identified in the sensitivity analysis as being highly sensitive. The results of this assessment are summarised in Table 7.
Table 7. Evidential uncertainty assessment of sensitive nodes

<table>
<thead>
<tr>
<th>Node</th>
<th>Evidential uncertainty rating</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety Behaviour</td>
<td>Moderate</td>
<td>The probability of insufficient safety behaviours is calculated based on the integration of GDF Physical Effect, Attention Management Capability and DLEAA Effect nodes. Its quantification uses HEART and expert judgment, but neither approach is capable of satisfactorily determining accurate probabilities for this node within this context.</td>
</tr>
<tr>
<td>Injurious Hazard Exposure</td>
<td>Significant</td>
<td>In the absence of data describing whether or not an individual has been exposed to a sufficient level of hazard to result in an injury, the quantification of this node is based on expert judgment.</td>
</tr>
<tr>
<td>Incident Type</td>
<td>Moderate</td>
<td>The incident types identified are the most common found in the accident data, but the figures from which their quantification was derived was based on limited data.</td>
</tr>
<tr>
<td>Attention Management</td>
<td>Significant</td>
<td>Very little is known about the specific effects of GDFs exposure on this capability. HEART was used to provide an estimate of its effect within the ‘Safety Behaviour’ node. While this method can capture failures in cognitive performance, it was not specifically designed to estimate failures in attention management capability.</td>
</tr>
<tr>
<td>Room Geometry, Space Allocation, Location of Areas, Accessibility / Circulation</td>
<td>Significant</td>
<td>The different states of the DLEAA input parameters node are set using qualitative ordinal categories based on subjective design assessment. The quantification of these nodes is based on judgment alone. HEART was subsequently used to provide an estimate of their effects within the ‘Safety Behaviour’ node. While this method can capture failures in cognitive performance, it was not specifically designed to estimate DLEAA-type effects.</td>
</tr>
<tr>
<td>Motion - Lateral Acceleration</td>
<td>Moderate</td>
<td>Although evidence exists of the effect of this node on the task performance, the ranges are set based on a limited amount of data generated from few relevant experiments.</td>
</tr>
</tbody>
</table>

Comparison of the results obtained from the models being in nomological proximity

Due to the limitation to the possibility of a high level comparison with the single model (Tomas et al, 1999) identified to have nomological proximity with the FGPIIM, a more detailed comparison of result is not possible.

Predictive validity summary

The sensitivity analyses have shown that the FGPIIM behaves as expected and generally reflects the industry data on personal injury outcomes; i.e. deaths are relatively rare in comparison with injuries (see D4.5, Table 13). When all the GDFs are set to their lowest effect (i.e. best) state, the probability of ‘No_Injury’ is at its highest. When all the GDFs are set to their worst state the probability of ‘Death’ (although very low) and the probability of ‘Injury’ both increase by 94% (when compared to the best case). This would suggest that the GDFs have an effect on the state of the outcome node.

The FGPIIM only estimates the probability of ‘no injury’, ‘injury’ and ‘death’ outcomes. The reference accident data does not allow a reliable distinction between minor and major injury, and hence greater fidelity in the FGPIIM could not be achieved.
While the sensitivity analysis shows the model operating as intended, the uncertainty and importance analyses reveal weaknesses within some of the most influential nodes. The primary source of weakness stems from a general lack of evidence that enables probabilities to be set for the elements within FGPIIM with confidence or permit validation against real world data.

Overall, the FGPIIM has a moderate level of predictive validity in its responsiveness to the manipulation of GDF states; however its accuracy in predicting personal injury to crew risk in absolute terms cannot be verified.

### 2.4 Construct related validity - fire risk model

#### 2.4.1 Background

The fire risk model (ref. FAROS D4.8 [28]) dealt with onboard fires in various functional spaces of cargo and passenger ships. The development was built upon related research in the previous EC funded projects such as SAFEDOR and FIREPROOF, as well as fire science literature and fire outbreak records onboard of various ships. As for the latter, Figure 2-13 illustrates the consulted data sources in terms of their volume and level of detail; the 150 records database was most detailed as it was based on public accident investigation (forensic) reports.

![Figure 2-13 Used sources of fire incident records](image)

In addition to the above material, we also utilised:

- Computer simulations (CFD was used) to understand physical behaviour of flammable oils transported in pressurised piping systems. Specifically, we were interested in determining the extent of spray (leak) when a pipe or other component develops a crack.

- Human error probability (HEP) databases and corresponding methods such as HEART were used to work out HEPs under given circumstances such as the effect of the global design factors, time pressure and others.

- Expert opinions were gathered through structured questionnaires about the time to failure and failure modes of system components in engine rooms. Specifically, chief and 2nd engineers were asked to provide the subjective estimates. The
collected data was then processed using formal techniques to converge at failure probability distributions.

Due to stochastic nature of fire outbreak and development [29], the probabilistic framework was also adopted to quantify probability of accident (i.e. ignition) and its consequences in terms of expected number of fatalities, i.e. potential loss of life (PLL). The consequence model was not addressed in FAROS and was merely adopted from the forerunner SAFEDOR [30]. Since then, the consequence model has been further developed and successfully applied in a number of commercial projects [31]. The latter fact gave us confidence in the consequence model, which we assumed to be validated through practical applications.

2.4.2 Translation related validity

Hence, the objective of this part of the report is to review the validity of the fire ignition model which was developed in the course of FAROS. The ignition model was developed for several onboard spaces as the fire space of origin is stochastic. The following onboard spaces were addressed, also see Table 8:

- Crew and passenger cabins.
- Large public spaces (e.g., restraints, lounge areas).
- Machinery spaces (e.g., engine room, generator room).
- Cargo holds.
- Other spaces (e.g. galley).

Table 8. Modelling and validation of ignition probabilities (based on [28])

<table>
<thead>
<tr>
<th>Space type</th>
<th>Ignition probability quantification approach</th>
<th>Validation aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crew and passenger cabins</td>
<td>The probability was expressed as a union of probabilities of ignition caused by smoking, electrical fault and arson. The marginal probabilities were inferred from available fire records, ignition properties of combustible material available in cabins, whereas human error probabilities – used to model carelessness during smoking – were taken from the HEART database.</td>
<td>The modelling approach was simple and completely transparent as it was based on classic probabilistic rules of combining conditional and marginal probabilities of events (i.e. the low of total probability). In the report, we also compared the ignition probabilities with statistic on fire accidents and explained the quantitative difference: the ignition probability is always greater than fire probability because only a fraction of ignitions leads to significant fires that get reported. As this ignition model was not based on the Bayesian Network, further validation excluded from this report.</td>
</tr>
<tr>
<td>Large public spaces</td>
<td>After the literature analysis, we concluded that the ignition probability equals to probability of electrical fault that may cause ignition. The latter was modelled using the Bayesian updating (statistical inference) of subjective probabilities. Thus, the modelling was based on statistics and classic methods of statistical inference. The human error was not included in the</td>
<td>We also compared the ignition probabilities with statistic on fire accidents and explained the quantitative difference: the ignition probability is always greater than fire probability because only a fraction of ignitions leads to significant fires that get reported. As this ignition model was not based on the Bayesian Network, further validation excluded from this report.</td>
</tr>
</tbody>
</table>
Although some historical data was also available it was only sufficient to assume the most common ignition scenario: contact between flammable oil and hot temperature surfaces. The ignition probability was expressed as a product of conditional and marginal probabilities. The conditional probabilities reflected physical properties of the flammable oils and hot surfaces, as well as the size of machinery space. Whereas the marginal probabilities for the flammable oil and hot temperature surfaces to become available were modelled using expert judgment and human error probabilities (HEP) taken from HEAR and NARA databases. Additionally, the human error probability was modelled as a Bayesian Network.

In the report we made the following assessment of validity:
- The modelling approach was simple and completely transparent as it was based on classic probabilistic rules of combining conditional and marginal probabilities of events (i.e. the low of total probability).
- The accuracy of modelled probabilities / reoccurrence intervals of flammable oil and hot temperature surfaces was confirmed with the ship operators participating in the project consortium.
- The predictability of the model was checked with the available statistics, explaining the reasons for deviations.

Because the model included a Bayesian Network, further arguments for its validity, or a lack thereof, are given in the following section.

The ignition probability was modelled by using the Bayesian Network. The main sources of information to support modelling assumptions were the existing safety regulations and expert opinions elicited within the company Brookes Bell LLP.

The model included the human error.

The probability was expressed as probabilities of ignition caused by electrical fault or failure of deep fat fryers, depending on the space purpose. The marginal probabilities were inferred from available fire records, whereas human error probabilities – used to model carelessness during the use of deep fat fryers – were modelled using human reliability assessment (HRA) technique CARA. The auto-/piloted ignition in the galley took into account the galley floor area and other design variables.

The model included the human error.

After further investigation, we found that the model does not sufficiently capture the ignition causality in cargo holds. On this basis the modelled was not applied in the project and hence is not addressed in this report as well.
2.4.2.1 Face validity

As described in deliverable D3.4 [33], the deck layout, equipment access and arrangements (DLEAA) affects human performance directly and indirectly. It is clear that DLEAA presents certain physical and cognitive demands upon the seafarer which they must be able to meet in order to perform a task. These may be physical, due to factors such as confined space or impaired accessibility, or cognitive and working memory demands, due to factors such as the distance and separation between functional areas.

As for the physical demands, DLEAA affects task performance, making specific tasks easier or harder to complete. Therefore the effect is relatively easy to grasp and quantify, assuming it manifests itself in physical obstacles that impede freedom of movement necessary for easy and therefore successful (supposedly) task execution. Examples of such physical obstructions are:

- Lack of space around the equipment under repair;
- Narrow pathways that require a mariner to slow down to avoid accidental and injurious contacts, especially when carrying heavy equipment;
- Obstructions on the way from one space to another (e.g., watertight doors, stairs and ladders), reducing the average walking speed and hence delaying execution of the task, which in turn may lead to time pressure.

According to the study by Case Western Reserve University [34], it is not necessarily the time pressure, but it is the perception of that time pressure that affects an individual. If one feels they don't have enough time to do something, it is going to affect them. The time pressure is proportional to the ratio between (1) the estimated, or perceived, time it
takes to complete a task at hand and (2) time required by external circumstances for task completion. The former might be fixed, for the number of crew members involved in the task, travel times between functional spaces, and time it takes to apply tools to fix a detected problem, for example, are more or less predetermined (e.g. the deck layout is fixed). In turn, the time required for task completion vary depending on different stages of operation, level of emergency etc. What is clear is that by reducing the estimated time it takes to complete a task at hand (i.e. when a person thinks he/she needs a short time to complete the task), the time pressure should potentially be reduced.

Therefore, unreasonable collocation of areas, i.e. segregation of functional spaces, resulting in longer travel times would contribute to time pressure. Analogically, insufficient space allocation around equipment to be repaired might impair time efficiency on certain tasks.

As far as detection, assessment and maintenance performance of tasks is concerned, it was assumed that, amongst other factors, it is also affected by

- Availability of sufficient working area in a given room. This relates to space allocation, see [33].
- Availability to maintain normal walking speed. This relates to colocation of areas, see [33].

The detection, assessment, and maintenance task performance were implemented as two separate nodes in a Bayesian network, affected by two corresponding nodes, as shown in Figure 2-14. It was assumed that the detection and assessment performance is affected by space allocation, i.e. sufficiency of space around equipment, and attention management capability (AMC) only (see deliverable D3.4 - [33] - for explanation). That is, collocation of areas, i.e. the distance between functional spaces, does not have any effect on success of detection and assessment. In turn, the maintenance task performance is affected by both characteristics of the deck layout, and also AMC. The understanding here is that the maintenance tasks involve not only local work around the equipment—hence there must be enough space around it—but also possible travelling along the deck (or across multiple decks) visiting different compartments (e.g., workshop, engine room). The effect of time pressure comes into play in the maintenance task performance.

Based on the above, we can provide the answer to the following questions.

**Q1:** Does the model structure (the number of nodes, node labels and arcs between them) look the same as the experts and/or literature predict?
Yes.

**Q2:** Is each node of the network discretized into sets that reflect expert knowledge?
Yes, as far as our design and operational expertise is concerned.

**Q3:** Are the parameters of each node similar to what the experts would expect?
Yes.
2.4.2.2 **Content validity**

Q1: Does the model structure contain all and only the factors and relationships relevant to the model output?
The model structure contains the main design information that is normally available at the concept design stage. However, other details may be included to reflect the differences between ship types and sizes.

Q2: Does each node of the network contain all and only the relevant states the node can possibly adopt? 
Yes.

Q3: Are the discrete states of the nodes dimensionally consistent? 
Yes.

Q4: Do the parameters of the input nodes and conditional probability tables (CPTs) reflect all the known possibilities from expert knowledge and domain literature? 
Yes.

2.4.3 **Criterion related validity**

2.4.3.1 **Concurrent validity**
This is not applicable because we found no model to compare with.

2.4.3.2 **Predictive validity**
The presented sub-model is very simple. Its predictive validity is directly dependent on the accuracy of the sub-model on the right hand side and the correct quantification of the deck layout by means of the two top nodes on the left. As for the latter, the probabilities were calculated by adopting the cumulative density function (CDF) of the exponential distribution, which is useful to model phenomena that involve exponential growth/decay with the constant average rate. The decay coefficients in the CDF were calculated by factoring in:

- Required volume per person in the room for successful completion of various work tasks. According to the HSE guidelines, it is 11 m³ per person. Note, MSC/Circ.834 does not specify exact values.

- Walking speed that is considered typical, normal or even optimal in terms of effort requirements. According to the IMO evacuation guidelines (MSC.1/Circ.1238), average speeds for female and male crew members on a flat terrain are within 0.93-1.55 and 1.11-1.85 m/s, respectively. In calculations, average values 1.24 and 1.48 m/s can be used.

On this basis, we assumed that the predictive accuracy was reasonable, although we did not have any data or model to compare with.
3 Summary and conclusion

3.1 Summary

In this report we demonstrate the use of a validation framework as proposed by [4], to the risk models developed in the course of FAROS project. Due to the inclusion of human element as a central part of risk models and lack of empirical data to benchmark the developed models, the adoption of framework stemming from social sciences seems fully justified. The framework offers a systematic way for reasoning about the quality of the risk models developed and helps to identify the areas that require further research to improve the overall performance of the risk models. The framework encapsulates several validity tests; however they do not “prove” that the model results are correct. Instead they indicate the extent to which the model is a plausible representation of the object of inquiry. This relates to adopted understanding of risk and the adopted risk perspective, where no reference is made to an underlying “true” risk. The results of validation tests are gathered in Table 9, which shows that the overall validity of the risk models evaluated here is between moderate and high. High face validity informs about good operationalization of the analyzed concept, meaning that the structure of the risk models is adequate.

Moderate content validity informs that there is room for improvement when it comes to the parametrization of the models. This stems mainly from the lack of data about the analyzed phenomena. In order to increase the score in this validation test, an extensive research is needed dedicated solely to the assessment of the effect of GDFs on performance of a crew member and resulting probability of personal injury, fire or collision and grounding.

Low - Moderate concurrent validity reflects the actual state of the art in the analyzed field. Due to lack of compatibility between the existing risk models and these developed in the course of FAROS, the direct comparison of these is not feasible. In some cases only elements of the models can be compared, but event then due to substantially different modelling paradigm behind the models the results of comparison are not satisfactory.

Moderate – High predictive validity shows that the risk models developed behave as expected when tested. Also the sensitivity-uncertainty assessment allowed defining the most critical elements of the models and remedial actions were taken to improve the overall performance of the model. These comprise the alternative hypothesis with respect to the most critical element. This high score supports the adopted modelling techniques, which make it possible to carry out an extensive tests pertaining to the predictive validity.

Table 9. The results of validation test for the risk models developed in the course of FAROS project (FGPIM, FCGRM, Fire risk model)

<table>
<thead>
<tr>
<th>Validity criteria</th>
<th>Risk model</th>
<th>FGPIM</th>
<th>FCGRM</th>
<th>Fire risk model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td></td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Content</td>
<td></td>
<td>Moderate</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Concurrent</td>
<td></td>
<td>Low-Moderate</td>
<td>Low-Moderate</td>
<td>NA</td>
</tr>
<tr>
<td>Predictive</td>
<td></td>
<td>Moderate</td>
<td>Moderate-High</td>
<td>Moderate</td>
</tr>
</tbody>
</table>
3.2 Contribution to FAROS objectives

Initially the purpose of the task described in this report was to cross-validate personal and societal (ship level) risk figures collected from the experimental data, with corresponding predications by risk models. However due to complexity of the topic analyzed the quantitative findings from the experiment conducted remain inconclusive. Therefore the outcome of the risk models cannot be compared with the experimental data.

Moreover, due to adopted constructivist risk perspective, where the risk is seen in wider sense that just a number, the validation of risk models shall by no means be limited to benchmarking with the historical data\(^5\). For that purpose, the wider validation framework was proposed and successfully applied to the risk models developed here. The framework allows risk models validation even in the lack of data, which was the case here. The validation exercise carried out demonstrates that the risk models developed are good representation of the objects of inquiry.

Therefore, the proposed validation framework clearly contributes to the project objectives and WP7 objectives.

3.3 Implementation

As the results of validation tests, it is justified to conclude, that the risk models developed are plausible enough to serve as a basis for further reflections, allowing for the differentiation between various ship designs. Thus the risk models developed can be used at the initial stage of ship design, as initially intended, allowing for relative comparison of various ship designs with respect to a predefine baseline design. However, it is not advised to apply risk models to the optimization process of ship design along with the utilitarian paradigm of decision making, especially when it comes to cost-benefit analysis. The main concern is about the proper reflection of costs related to a risk level calculated by the models. Since the risk models are of high uncertainty already, expanding them by another dimension of uncertainty related to the costs is not advisable. Rather the user should seek the ways to reduce the uncertainty, and arrive at convincing conclusions. Comparing various ship designs to a common baseline, and use the relative risk reduction values, instead of absolute values, is one way of uncertainty reduction.

In that model, the risk models may be used by naval architects, vessel designers, and vessel system designers as intended, provided access to HF expertise is available to assist with application and interpretation. It is important to recognise the relevance of human factors input during its eventual application. HF provides the understanding of the complexities of human behaviour in operational settings, its interdependencies and interactions.

\(^5\) Benchmarking with historical data on ships accident is very problematic here, and even if conducted it remains inconclusive. The reason is that the developed risk models are very specific; they tend to measure the effect of specific manipulation in ship design on the probability of certain type of accidents. The statistics however are very generic. They rarely inform about the causes of an accident (frequently the cause is classified either as a human or technical error, which is not very informative). Moreover the historical data are affected by the issue of accident and incident underreporting, which if accounted for, will change the existing accident frequency obtained from the statistics. This means that we may expect the results of such a benchmarking exercise being rather misleading than informative.
4 Bibliography and References

5 Indexes

5.1 Index of Tables
Table 1: The qualitative assessment of evidential uncertainty for models assessing the probability of an accident, [25] ........................................................................................................ 28
Table 2: The qualitative assessment of model parameters importance for models assessing the probability of an accident [25] ........................................................................................................ 29
Table 3: The probabilities of collision and grounding obtained in the course of extreme conditions tests .................................................................................................................................................. 30
Table 4: The results of comparison of two nomologically proximate collision risk models. ........................................................................................................................................................................ 31
Table 5: Node groupings by sensitivity ................................................................................................................................. 38
Table 6: Uncertainty assessment rating criteria .......................................................................................................................... 38
Table 7: Evidential uncertainty assessment of sensitive nodes .................................................................................................. 39
Table 8: Modelling and validation of ignition probabilities (based on [28]) ......................................................................................... 41
Table 9: The results of validation test for FGPIM and FCGRM .................................................................................................................. 46

5.2 Index of Figures
Figure 1-1 The structure of validity framework adopted here ............................................................................................................ 10
Figure 2-1 Causal chain describing the relationship between crew GDF exposure and unwanted outcomes ........................................................................................................ 15
Figure 2-2 Supporting data for links in the causal chain, [13] ................................................................................................................ 18
Figure 2-3 DNV collision model .................................................................................................................................................... 25
Figure 2-4 FAROS collision risk model .............................................................................................................................................. 26
Figure 2-5 The results of extreme conditions tests performed on collision and grounding risk models. ................................................................................................................................. 30
Figure 2-6 The non-dimensional results of comparison of two nomologically close collision risk models: FCGRM and DNV model. ........................................................................................................ 31
Figure 2-7 Structure of the FAROS Generic Personal Injury Model (FGPIM) .................................................................................. 32
Figure 2-8 The FAROS Generic Personal Injury Model (FGPIM) .................................................................................................... 33
Figure 2-9 The hypothetical Tomas’s model ........................................................................................................................................... 36
Figure 2-10 Comparison of FGPIM and Tomas’s models ......................................................................................................................... 36
Figure 2-11 Comparison of FGPIM and Tomas’s models .......................................................................................................................... 40
Figure 2-12 Structure of the ignition model (only main directly influencing factors are shown) ..................................................................................................................................................... 16
Figure 2-13 The assumed causal chain ............................................................................................................................................... 17
Figure 2-14 Bayesian network for estimating probability of sufficient detection, assessment and maintenance task performance ........................................................................................................... 43

5.3 Abbreviations
GDF – Global design factors, such as ship motion, noises and vibrations.
HRA – Human Reliability Analysis
AMT – The Attention Management Theory
HPM – Human Performance Model
BBN – Bayesian Belief Networks
FCGRM – FAROS Collision and Grounding Risk Models – the models provide the
probability of collision or grounding respectively, to a ship given set of GDFs. The output can be further used to assess the societal or environmental risk levels for a fleet of ships or a given sea area.

FGPIM – FAROS Generic Personal Injury Model provides the probability of personal injury, as the result of exposure to an injurious hazard given a set of GDFs. Distinction is made only between three levels of personal injury outcome: no injury, injury and fatality. The output of this model can be further used to assess individual risk levels for a given ship.