Bayesian aggregation of expert judgment data for quantification of human failure probabilities for radiotherapy

L. Podofillini, D. Pandya, F. Emert, A.J. Lomax & V.N. Dang
Paul Scherrer Institute, Villigen, Switzerland
G. Sansavini
Polytechnic of Zürich, ETH, Switzerland

ABSTRACT: The paper deals with the quantification of probabilities for human failures in the radiotherapy domain. The probabilities are used as input for the development of a Human Reliability Analysis (HRA) method specific for radiotherapy. Quantification is based on expert judgment, in view of the lack of relevant data. A Bayesian aggregation model is used to aggregate the judgments collected during elicitation sessions with domain experts. A qualitative scale is first used; then the judgments are interpreted as information on the order of magnitude of the error likelihood and aggregated under the Bayesian scheme. Besides for the specific domain of interest, this work is relevant for novel HRA applications outside typical domains, for which the need to incorporate expert judgment in traceable and defendable ways is key.

1 INTRODUCTION

Human failures are important contributors to near misses, incidents, and accidents in radiotherapy (WHO 2008), as in many other domains. Efforts are undertaken to systematically address the potential for failures and continuously improve the patient treatment process, e.g. Huq et al. (2016). In this context, the Risk and Human Reliability research group at the Paul Scherrer Institute (Switzerland), in collaboration with the institute’s Center for Proton Therapy, is developing a method to support Human Reliability Analysis (HRA), specific for external beam radiotherapy. Previous work by the authors identified the personnel tasks critical to patient safety and possibly influencing factors (Pandya et al. 2017). Current work is addressing the quantification of the corresponding human failure probabilities.

In particular, the present paper focuses on the quantification of the failure probabilities for representative tasks, given a set of Performance Influencing Factors (PIFs). Given the shortage of directly usable experience data, the quantification resorts to expert judgment. The paper presents the application of a Bayesian aggregation model (Podofillini and Dang, 2013) to the judgments collected during elicitation sessions with domain experts. To avoid direct elicitation of probability values, the experts are asked to provide their judgments on a qualitative scale. The judgments are then interpreted as information on the order of magnitude of the error likelihood and aggregated under the Bayesian scheme. The paper presents the results of the aggregation. The application shows the ability of the aggregation approach to formally represent the variability of the experts’ estimates.

Besides for the radiotherapy domain, the work presented in this paper is relevant for the various efforts recently done to extend HRA methods for application beyond their most typical applications, i.e. nuclear power plant operation. Lack of relevant data is a major issue for these novel applications (Bye et al. 2017, NASA 2012, Gibson 2012, Mkrtchyan et al. 2015, NUREG 2016) and methods to elicit expert judgment in a formal and defendable way are needed along with specific data collection initiatives.

The paper is organized as follows. The next Section provides the background on the HRA method under development, for which probability values are sought for in this paper. Section 3 presents the design of the elicitation sessions and the concepts underlying the Bayesian approach for processing and aggregation of the judgments. Section 4 presents the application to two Decision Trees part of the framework of the HRA method under development. Concluding remarks close the paper.

2 BACKGROUND INFORMATION

The framework for the HRA method consists of eighteen decision trees, one for each failure mode
corresponding to a different Generic Task Type (GTT, Table 1). The concept behind the GTTs is taken from the Human Error Assessment and Reduction Technique (HEART, Williams 2017), and is intended to define a set of task types, each with similar characteristics as it relates to the factors influencing performance and to the corresponding failure probabilities. The definition of the GTTs and of the influencing factors is based on GTTs- Performance influencing Factors (PIFs) structures developed in (Pandya et al. 2017); these structures link each GTT to the set of PIFs that influence the failure probability. These structures have been developed via a systematic and traceable process which, for each GTT, progressively identifies the involved cognitive functions, their failure modes and causes, failure mechanisms and PIFs.

The DT framework is well suited to represent the cause-based influences on failures identified by the GTT-PIF structures: the DTs identify the causes possibly leading to the GTT failure; in a similar way as done in other HRA methods (e.g. NUREG 2016, Moieni et al. 1994), each decision tree addresses a failure mode, with branching points representing the effects of PIFs. Two examples of DTs are presented in Figure 1. The decision trees develop from eight branching points, e.g. “Problematic interface”, “Information content unclear”, “Low vigilance due to expectations”. As shown in Figure 1, each DT includes a subset of the eight branching points, three or four in most cases. The same branching point heading may appear across different DTs, e.g. “Problematic interface” in Figure 1; however, the influence of the branch on the failure probability may not necessarily be the same. This aspect will be returned to in the result Section 4. Each branch point is specified in terms of negative conditions: if any of the negative conditions is verified, then the lower branch applies. Example negative conditions are given in Table 2. The presentation of the development of the DTs from the GTT-PIF framework and of the negative conditions for each branching point is outside the scope of the present paper and will be presented in a separate publication (Pandya et al., working paper).

To assess the failure probability of a specific radiotherapy task, an analyst would have to select the applicable DTs based on the relevant type of task and failure mode. Then, for each branching point, the analyst would have to select the appropriate branch based on the negative conditions proposed for each branch, in a similar way as done with other HRA methods involving DTs, e.g. NUREG (2016), Moieni et al. (1994).

The present paper focuses on the quantification of the DTs, i.e. on the assessment of the failure probabilities in correspondence of each path defined by the combination of the branching points.

### Table 1. Set of Generic Task Types (GTTs) and corresponding failure modes identified in Pandya et al. (2017). DTs are developed for each failure mode of the GTTs.

<table>
<thead>
<tr>
<th>#</th>
<th>GTT</th>
<th>Failure mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identification of patient or patient related items</td>
<td>Patient information incorrectly matched Identification check not performed (decision based) Failure to execute desired action</td>
</tr>
<tr>
<td>2</td>
<td>Quality Check</td>
<td>Deviation from requirement not recognized Inappropriate understanding of underlying principles Check not performed (decision based) Execute desired action incorrectly Failure to execute desired action</td>
</tr>
<tr>
<td>3</td>
<td>Complex interaction with software or tool</td>
<td>Coordination failure Misinterpretation of data Execute desired action incorrectly Mismatch or inconsistency not recognized</td>
</tr>
<tr>
<td>4</td>
<td>Simple interaction with software or tool</td>
<td>Execute desired action incorrectly Failure to execute desired action</td>
</tr>
<tr>
<td>5</td>
<td>Iterative determination of optimum parameters</td>
<td>Misinterpretation of information Inappropriate decision on strategy selection</td>
</tr>
<tr>
<td>6</td>
<td>Verbal communication</td>
<td>Communication failure Not communicated (decision based)</td>
</tr>
</tbody>
</table>

### Section 3. Expert Judgment Elicitation and Aggregation Approach

#### 3.1 Expert judgment elicitation

As mentioned in the Introduction, due to the lack of relevant data, quantification is made via expert judgment. In particular, the expert elicitation sessions were designed with two aims. First, to support the identification of the negative conditions underlying each branch point. Second, to assess the impact of each branching point on the failure probability. Only the effects of single branch points were addressed (i.e. determining failure probabilities 1, 2, and 4 in Figure 1, top part). The combination effects will be addressed in future work.
Six failure scenarios were developed for the elicitation. Table 3 gives two examples. The idea is to elicit the impact of the branching point on these failure scenarios, which would then be representative of the overall GTT. The selection aimed at addressing the largest set of GTT failure modes, as well as prioritizing failure scenarios with the most critical consequences on patient safety. As shown for the examples in Table 3, each failure scenario is associated to a different GTT. Indeed as again shown in Table 3 and by the negative conditions in Table 2, the elicitation of the branching point impact was conducted on specific tasks and situations. This has been made to help experts to contextualize their judgments to the real tasks they perform and link their assessments to the daily experience. Alternatively, judgments may have been elicited directly for the GTTs and branching point categories. The former approach was chosen to avoid that experts would need to deal with abstract categories such as GTTs and the branch point labels. The focus of this paper is on quantification of the failure probabilities at each tree branch (only single branch effects elicited).

Table 2. Examples of negative conditions for two branching points in two different DTs.

<table>
<thead>
<tr>
<th>DT</th>
<th>Branch point</th>
<th>Negative conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTT: Identification of patient or patient related items; Failure mode: patient information incorrectly matched</td>
<td>Problematic interface</td>
<td>The written values look alike (e.g. 111, 117)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The value on the label or file not easily readable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>There is no ID number on the patient item</td>
</tr>
<tr>
<td>GTT: Complex interaction with software or tool; Failure mode: Misinterpretation of data</td>
<td>Lack of training or experience</td>
<td>Lack of familiarity with the tumor case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lack of training or experience on treating special tumor locations (e.g. close to multiple artefacts)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lack of experience or training to distinguish healthy and tumor tissues</td>
</tr>
</tbody>
</table>
Indeed, the elicitation addressed directly only part of the DTs required for quantification, i.e. six out of the eighteen from Table 1. However, some branching points may be thought of having very similar impact across different DTs; therefore the results from the elicitation for one DT may be used for others. It was assessed that the selected tasks may allow to quantify about two thirds of the whole set of branching points (recall only single branch point effects are considered here). Indeed future work may address the quantification of the remaining DTs and develop an approach to address multiple branch points as well.

Twelve experts were interviewed: medical physicists, medical doctors, dosimetrists and radiation technologists. Each expert dealt with tasks part of his/her daily job. Three tasks were elicited at most per expert. Each expert took part in the exercise alone. For each of the assigned tasks and each of the negative conditions corresponding to the branching points, the experts were asked to assess the impact of the negative condition on the failure probability when performing the task. The impact is elicited on a qualitative scale (Table 4), to avoid the known shortcomings of directly eliciting probability values, see e.g. Meyer and Booker (2001), Tversky and Kahneman (1974).

### 3.2 Aggregation of expert assessments

The approach to process the expert assessments has been as follows. The qualitative scale in Table 4 is first anchored to quantitative values, as shown in Table 5. The basis for the anchoring is the scale

<table>
<thead>
<tr>
<th>Impact</th>
<th>Descriptor</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low impact</td>
<td>Failure is not expected to happen, although I see how it could happen.</td>
<td>Given the negative condition, the desired task is still so easy that it is inconceivable that any personnel would fail if they were to experience this condition.</td>
</tr>
<tr>
<td>Moderate impact</td>
<td>Failures happen occasionally/sometimes with such conditions</td>
<td>Given the negative condition, the desired task becomes moderately difficult that it is possible so that personnel would occasionally/sometimes fail if they were to experience this condition.</td>
</tr>
<tr>
<td>High impact</td>
<td>Failures happen often with such conditions</td>
<td>Given the negative condition, the desired task becomes highly difficult that is expected so that personnel would often fail if they were to experience this condition.</td>
</tr>
<tr>
<td>Extreme impact</td>
<td>Failure is almost unavoidable</td>
<td>Failure is almost unavoidable. Almost all personnel would not be able to perform the desired task.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Impact</th>
<th>Order of magnitude of failure probability</th>
<th>Table 5. Anchoring of the qualitative impact scale to probability values (adapted from NUREG 2007).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1e-3</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>1e-2</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>1e-1</td>
<td></td>
</tr>
<tr>
<td>Extreme</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Example of failure situations used to elicit the impact of the branch points on the failure probability.

<table>
<thead>
<tr>
<th>Failure situations</th>
<th>Failure mode</th>
<th>Generic task type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to identify correct ID from control document on the bite-block, couch or file etc. such that incorrect item is picked up</td>
<td>Patient information incorrectly matched</td>
<td>Identification of patient or patient related items</td>
</tr>
<tr>
<td>Draw suboptimal (incorrect or incomplete) contours around volumes of interest for every slice due to misunderstanding of the data</td>
<td>Misinterpretation of data</td>
<td>Complex interaction with software or tool</td>
</tr>
</tbody>
</table>
presented as part of the ATHEANA human reliability analysis method (NUREG 2007). Note the values on the scale are to be interpreted as reference orders of magnitude for the failure probability value. The value representative for low impact (1e-3) is confirmed in the studies by Wahi et al. (2008) and Salinas et al. (2013), from which it can be inferred that nominal error rates in patient identification and data entry in healthcare lie around 1e-3 and 3e-3. These are interpreted as lower bounds for error probabilities for the sector. Low impact of the branching point is not expected to change the order of magnitude of the probability so that the reference lower bound value still remains in the same order of magnitude.

The scale allows converting each assessment by the experts into a statement on the order of magnitude where the probability value would lie. It is interpreted as evidence of the relevant order of magnitude, and used to update the belief on that quantity in a Bayesian framework.

The process for aggregation of the judgments comprises two steps. First, for each negative condition, the judgments by the experts are aggregated: a distribution of the applicable probability for each condition is obtained. Then, these distributions are themselves aggregated into the final distribution of the corresponding branch. The aggregation approach is based on the Bayesian model presented in Podofillini & Dang (2013). The model represents the human error probability as an inherently variable quantity, resulting from the inherent variability of people performance as well as of the specific manifestations of the type of tasks and of the influencing factors. More specifically, the combination of GTTs and branch point conditions envelop specific tasks and specific performance conditions that are assumed to be characterized by inherently different failure probability values. The elicitation carried out in this work addressed specific manifestations of the combination (see Table 3 and Table 2): the Bayesian model is intended to consider the expert assessments on these manifestations (a specific task affected by specific negative performance conditions) and determine the original variability distribution of interest. Mathematically, the failure probability is assumed to be lognormally distributed, with unknown median to be determined based on the expert input. The error factor (square root of 95th and 5th percentile) is assumed to be known, equal to 3. The latter assumption of known error factor is not a requirement of the approach, but largely simplifies the calculations and decreases the amount of data required to be elicited. It is indeed a typically used and accepted value in HRA. The prior distribution of the median is assumed uniform for the four orders of magnitude in Table 5 (all impact levels are equally likely).

For the first part of the aggregation process, the expert assessments are used to update the degree of belief on the correct order of magnitude for the median of the probability distribution. The model requires as well assumptions on the confidence that the experts would be able to provide the correct value of the probability. The confidence is expressed in terms of a conditional probability that, given the real order of magnitude of the probability is one of the four in Table 5, the experts would assess the correct one or be off by one or more orders of magnitude. This conditional probability can be defined to model biases and dependence across experts, indeed provided that adequate information on the distribution is available (these are not modeled in the present work). In this work, it is assumed that experts have about 80% probability to provide the correct order of magnitude, 10% of being one order of magnitude off, 5% of being two or more orders of magnitude off. The exact values of these probabilities depend on the position of the interval with respect to the lower and upper bounds to have them normalized to a probability distribution. These values have been assumed by the authors of the paper; they appear to represent reasonable assumptions on the ability of the experts to provide correct estimates in this context. It is anyway important to mention that as more than a few experts are available (say five or more), the specific assumptions on the confidence to each expert do not play a significant role anymore in the final probability distribution. The output of this step is a distribution of the degree of belief on which of the levels in Table 5 represents the real value of the median of the probability distribution, for each negative condition possibly affecting each branch point.

The second part of the aggregation entails combining the degrees of belief obtained for each negative condition underlying each branch points. As the negative conditions are assumed equally likely, the final distribution is simply obtained as the average distribution across the negative conditions. In particular, for each of the levels in Table 5, the final degree of belief is the average degree of belief across the corresponding negative conditions. Applications of the approach will be presented in the next section.

4 AGGREGATION OF EXPERT ASSESSMENT: RESULTS AND DISCUSSION

This paper presents the result obtained for two DTs:
• GTT “Identification of patient and patient related items”, failure mode “Patient information incorrectly matched”;
• GTT “GTT: Complex interaction with software or tool”, failure mode “Misinterpretation of data”.

These are the two DTs shown in Figure 1. The results obtained from the whole elicitation are planned to be presented in Pandya et al. (Working paper).

The first part of this Section presents an overview of the aggregated results for the two DTs. The aim is to discuss how the quantitative results relate to the justification provided by the experts on their assessments. In other words, the goal is to check if the different values of error probability reflect in corresponding differences in the assessments by the experts. The second part of the Section provides details on how the expert assessments are aggregated.

Figure 2 shows the aggregated results from the expert assessments. The largest impact on failure probability corresponds to the branch point “Lack of training or experience” branching point affecting GTT “Complex interaction with software/tool”. This is a complex task, related to defining an optimal therapy plan and requiring knowledge and expertise. As shown by the expert assessments, the influence of inadequacies in this respect can have high impact on the failure probability. Indeed, the resulting median probability is around 0.01, corresponding to the “high” impact level on the adopted scale. Branching point “Time pressure” was also assessed among the most influencing ones: it was felt that the need to complete the task with urgency would highly impact the quality of the therapy plan. On the other hand, two branching points were assessed to have generally low impact. In particular, interface issues (branching point “Problematic interface”) were not felt to affect much the performance when identifying patients: identification of patients is made with diverse means; besides checking the patient ID, identification is checked verbally (calling patient name) and by the patient picture. Additionally the interface to deal with is extremely simple so that there is little possibility for confusion. Also, the complexity, in shape, size and location of the tumor was not felt to increase much the probability of errors in the development of the therapy plan. Typically, complex tumor cases are discussed in larger groups and the treatment details are thoroughly defined.

It is interesting to see in Figure 2 how the same type of branching point may affect GTTs differently. For example, the assessed probability for “Problematic interface” affecting GTT “Complex interaction with software/tool” is more than one order of magnitude larger than when affecting GTT “Identification of patient or patient-related tools”. Again, this is the result of the expert opinions on the importance of the respective influences. The reasoning underlying the low impact according to the experts of the branch point on the latter GTT has already been discussed. On the other hand, the former GTT involves complex interactions with multiple software interfaces, dialog boxes, figures, etc: the impact of interface issues for this task was considered to have important effects on the failure probability.

The length of the error bars reflects differences in the expert assessments both for each negative condition and across the different conditions. The larger the bar, the larger the differences. This aspect will be returned to later in this Section.

Figures 3 and 4 show how the expert assessments are progressively processed to obtain the final probability distributions presented in Figure 2. The left side of the figures gives the assessments of the experts provided on the scale for each of the negative conditions. The middle shows the distribution results aggregated across the experts for each negative condition. The right side gives the final distribution, aggregating across the conditions. Specifically, Figure 3 addresses the effect of branching point “Problematic interface” on GTT
“Identification of patient or patient-related items”.

It is interesting to see in the Figure how different assessments from the experts result in different distributions. For the first negative conditions, three experts provided the assessments of “low” (thus corresponding to an error probability of about 0.001) and two of “moderate”. Correspondingly, the aggregated distribution in the middle of Figure 3 presents larger degree of belief for the latter level on the scale compared to the former. Degrees of belief for the other levels are in practice negligible. For the second condition, there is strong agreement for low impact: the peak in the degree of belief for the latter value is accordingly higher than in the previous case.

A very different situation is present for the last condition, where the three experts provided three different assessments. The effect to spread the degree of belief for the latter condition is evident.

Another interesting case is presented in Figure 4, related to the effect of “Lack of adequate training or experience” for the GTT “Complex interaction with software/tool”. There is general consistency across the expert on the effect of each condition, as reflected in the distributions in the middle part of the figure. The large span in the expert assessments, from “low” to “extreme”, results then in the large spread in the final aggregated distribution on the right.

![Figure 3](image1)

**Figure 3.** Processing and aggregation of judgments (GTT “Identification of patient and patient related items”, failure mode “Patient information incorrectly matched”): Left: judgements from experts, Middle: expert-aggregated posterior distribution of median HEP for each condition, Right: posterior probability distribution of median HEP for the branch point.

![Figure 4](image2)

**Figure 4.** Processing and aggregation of judgments (GTT: “Complex interaction with software or tool”, failure mode “Misinterpretation of data”) Left: judgements from experts, Middle: expert-aggregated posterior distribution of median HEP for each condition, Right: posterior probability distribution of median HEP for the branch point.
5 CONCLUSIONS

The paper has presented the work performed to quantify the human failure probabilities to be used as input to a novel HRA method. Elicitation sessions were designed, with the following two main features. First, information on probabilities is asked to experts on a qualitative scale, with the goal of getting evidence on the order of magnitude for the probability. Second, specific situations are presented to the expert, i.e. specific failure scenarios influences by specific negative conditions. The latter feature was incorporated to avoid that the expert deal with abstract categories such as tasks types and influencing factors.

The expert statements are processed and aggregated to determine degrees of belief on the correct values of the failure probability. The latter is assumed as an inherently variable quantity so that the main parameter of its distribution is the subject of the elicitation. The Bayesian model used to aggregate the assessments was found to represent well the differences in the experts statements, providing a credible approach to process the expert input.

As a next step of the work, comparison of the obtained values with values from existing HRA methods is envisioned. Indeed, although HRA methods are sector-specific, some of the underlying data can be thought of being general, e.g. data regarding dealing with indicators, simple execution tasks. This comparison may provide some validation to the elicitation process.

With broader perspective, future work will apply the developed HRA method to hypothetical accident scenarios at the institute’s Center for Proton Therapy.

ACKNOWLEDGMENT

This work is funded by CROSS (a PSI inter-departmental funding initiative) and the Paul Scherrer Institute’s Center for Proton Therapy. The authors would also like to thank the personnel working at PSI’s CPT for their cooperation and support. The authors acknowledge the support from the Future Resilient Systems program at the Singapore-ETH Centre, established between the Swiss Federal Institute of Technology in Zurich (ETH Zurich) and Singapore’s National Research Foundation (FI 37007401).

REFERENCES
