Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

POTENTIAL AND CHALLENGES OF ASSURANCE CASES FOR SIMULATION VALIDATION

Pia Wilsdorf[1], Steffen Zschaler[2], Fiete Haack[1], and Adelinde M. Uhrmacher[1]

[1]Institute for Visual and Analytic Computing, University of Rostock, Rostock, GERMANY
[2]Department of Informatics, King's College London, London, UK

ABSTRACT

Simulation studies require thorough validation to ensure model accuracy, reliability, and credibility. While validation typically focuses on the simulation model itself, additional artifacts also influence study outcomes. Conceptual models, comprising research questions, requirements, inputs and outputs, model content, assumptions, and simplifications, provide context information for interpreting results and assessing model suitability. Validating other simulation artifacts for their fitness-for-purpose is complex, necessitating structured arguments to increase confidence. This paper explores when and how validation arguments should be constructed throughout the modeling and simulation lifecycle. Drawing on concepts from safety assurance cases, it defines key claims for the various artifacts, discusses validation arguments from different perspectives – including process, product, people, and project – and illustrates them through a computational biology case study. We conclude with a discussion of the suitability of such structured arguments for the comprehensive validation of simulation studies.

1 INTRODUCTION

In simulation studies, validation is crucial in ensuring the accuracy, reliability, and credibility of the simulation model and the achieved results. Typically, validation focuses on the simulation model, i.e., whether the right model was built for the system and answering the questions of interest (Balci 1997). However, additional artifacts play a role in conducting a simulation study, thus influencing its outcome. Robinson (2008a) introduced conceptual models for moving from a problem situation through model requirements to defining what should be modeled. In this process, the conceptual model becomes a collection of constructs and knowledge defined in various mostly non-formal forms that guide the simulation study and, in addition, is essential in interpreting the final results (Fujimoto et al. 2017). The conceptual model comprises the research questions, requirements, inputs and outputs, model content (characterized by a particular scope and level of detail (Stachowiak 1973)), various assumptions, and simplifications. This information provided by the conceptual model is crucial in determining whether the simulation model is suitable for answering the questions of interest about the system and thus might provide arguments for or against its validity (Uhrmacher et al. 2024). For example, in the artifact-based workflow by Ruscheinski et al. (2019), the validation of the simulation model implies that the fulfillment of all requirements needs to be checked. If the requirements refer to the expected behavior of the simulation model, simulation experiments are generated and executed to check the requirement automatically. At the same time, provenance information is collected (Ruscheinski et al. 2019) that can later be used to show that the requirement holds, as demonstrated by the associated simulation experiments.

In addition to the simulation model, the other artifacts and their role in the simulation study might equally be subject to being questioned. Osman Balci suggested applying validation, verification, and accreditation equally to the diverse modeling and simulation life cycle artifacts and adopting methods from software engineering to increase the quality of simulation studies (Balci 2012).

Validating whether a simulation is suitable for answering the questions of interest is, thus, a complex, multi-dimensional challenge. It is insufficient to run a simulation experiment, provide a test suite, or

have well-documented requirements. In fact, as also argued by Stepney and Polack (2018), because of the complexity of the systems being modeled (which is what makes simulation necessary), validity is contingent: we cannot demonstrate full validity, but we can provide arguments to reduce our uncertainty about the validity of a given simulation model and increase our confidence in its *fitness for purpose*. Alden (2012) proposes the use of structured arguments – as initially introduced for the capturing of safety assurance cases (Assurance Case Working Group (ACWG) 2021; Object Management Group (OMG) 2021) – to capture arguments about the fitness for purpose of a simulation model; this idea was later extended in Zschaler and Polack (2023) proposing the use of formal modeling languages to capture different aspects of simulation models, see also (Uhrmacher, Frazier, Hahnle, Klugl, Lorig, Ludäscher, et al. 2024), and their fitness-for-purpose arguments.

In this paper, we explore how assurance cases using structured arguments – inspired from safety-critical systems – can be constructed for the various artifacts of a simulation study. We define *assurance patterns,* consisting of key claims along with evidence and arguments needed to support them and instantiate them as *assurance cases* for artifacts from a simulation case study. This will provide a better foundation for understanding to what degree ideas about structured-argument models from safety assurance can benefit the engineering of simulation models and the conduct of simulation studies and where they are facing challenges. In applying structured arguments, it will not only be the question of whether requirements – for example, to reproduce various wet-lab data – were used successfully to calibrate the simulation model and how so (Haack et al. 2020), but whether stating these requirements was well-founded and why.

2 ASSURANCE CASES

Many engineered systems are safety critical – meaning that their "failure could result in loss of life, significant property damage, or damage to the environment" (Knight 2002). For safety-critical systems, it is, therefore, particularly important to clearly document what has been done to ensure they will not fail or to limit the impact of any potential system failure. To manage the complex arguments required for this, the safety-critical systems community has developed techniques for systematically capturing and managing safety assurance cases as *structured arguments*.

Two major modeling approaches for expressing such structured arguments exist: The Goal Structuring Notation (GSN) (Assurance Case Working Group (ACWG) 2021) was originally developed at the University of York and is now a community standard. Generalizing from GSN, the Object Management Group (OMG) (2021) developed the Structured Assurance Case Metamodel (SACM) to provide an industry standard. Wei et al. (2019) provides a good overview of the two standards, their relationship, and the tool support available.

GSN and SACM both build on Toulmin's (1958) layout of arguments, but use slightly different terminology for the fundamental building blocks of structured arguments. We will talk about a) *claims,* statements asserting properties or qualities of the system; b) *evidence,* supporting information, data, or analysis that substantiates the claims; and c) *arguments* (or *strategies*), the reasoning demonstrating why the evidence supports the claims (and which may hierarchically break down claims into sub-claims before eventually providing concrete evidence). In addition, assurance cases may provide context information, for example, justifying why a particular argumentation strategy is applicable.

In this paper, we will use the GSN notation to capture structured arguments for assurance cases for the validity of simulation artifacts. Solid rectangles represent claims, parallelograms represent strategies, and rectangles with rounded corners (we use this notation to save space over the circle notation used by GSN) represent evidence. Open diamonds indicate parts of an argument that have not been further developed.

Assurance patterns (Kelly and McDermid 1997) capture recurring argument structures in assurance cases. They are, in effect, parametrized structured arguments meant to be instantiated for specific systems. In discussing the different simulation-study artifacts, we will first provide typical assurance *patterns* for each artifact before illustrating them by instantiating some patterns as concrete assurance *cases* for a case study.

3 ASSURANCE FOR SIMULATION STUDY ARTIFACTS

This section illustrates the use of assurance cases to validate the different artifacts involved in a simulation study. How we approach answering the question "Did we choose the right artifact?" will depend on the type of artifacts and how they are specified. In the following, we will define assurance patterns (AP) for the conceptual model, the simulation model, and the simulation experiments. Each pattern will consist of claims, arguments, and evidence to be provided to give substance to the artifact's validity. We illustrate our approach based on a simulation study of endocytosis in the Wnt signaling pathway (Haack et al. 2020), which we introduce briefly.

Endocytosis is a cell-biological process in which membrane-integral receptors are internalized into the cell. Different mechanisms can mediate this process and may have different functional implications for the cell. The two most important internalization mechanisms considered in the simulation case study are clathrin-mediated endocytosis (CME) and caveolin-mediated/clathrin-independent endocytosis (CIE). In the case of the WNT signaling pathway, a central signaling pathway in development and cancer, internalization of the key receptor LRP6 plays a central role as it either hampers or promotes the signal transduction into the cell depending on the internalization mechanism employed. However, its exact regulation is still under debate, and different, even conflicting, theories exist about which internalization pathway is employed under which conditions. The simulation study aims to resolve this ambiguity.

3.1 Conceptual Model

We adopt the definition by Robinson (2008b), Balci (2012) and Fujimoto et al. (2017) and interpret the conceptual model as a collection of different artifacts involved in building and analyzing a simulation model.

3.1.1 Research Questions

Usually, a simulation study starts with a specific research question. However, new research questions may arise during a simulation study. The research question often refers to understanding the mechanisms behind some phenomenon observed in the real world, making predictions, or optimizing some product or process (Cellier 1991).

The precise formulation of the research question is crucial, as further artifacts of the simulation study will also be validated in relation to it (see the following subsections). The validity of the research questions cannot be assessed using formal verification, tests, or experiments. We define two assurance patterns that shall support the validity of a research question:

AP1. **Claim:** *the research question is relevant.* Different types of **arguments** can be provided in support of this claim (*cf.* also Balci's notion of the "four Ps" (Balci 2012), especially 'product', 'people', and 'process'), for example: a) by reference to the state-of-knowledge/state-of-the-art in the specific scientific domain, **arguing** that current knowledge indicates the research question is meaningful to ask and that it has not yet been answered; b) by reference to people, **arguing** that the research question has been defined by (or defined in collaboration with) appropriate domain experts and other stakeholders; or c) by reference to a process, e.g., **arguing** that a guideline or standard for identifying research questions in the domain was used (Goldschmidt and Matthews 2022). Supporting **evidence** will refer to the respective sources, e.g., publications, interview scripts, or standards.

AP2. **Claim:** *the research question is amenable to simulation.* Here, we may a) **argue** that similar questions have been answered by simulation studies, with the **evidence** of exemplary scientific literature; or b) **argue** about properties of the research question, e.g., that it pertains to dynamic processes, predicting or understanding complex systems, and that hypotheses about the mechanisms are available. **Evidence** might be given by references to related simulation models or to a causal, qualitative model (which might be part of the conceptual model; see below).

The **case study** investigates the research question (see Fig. 1): which of the (in different theories) suggested internalization mechanisms are most likely to mediate LRP6 receptor internalization in WNT

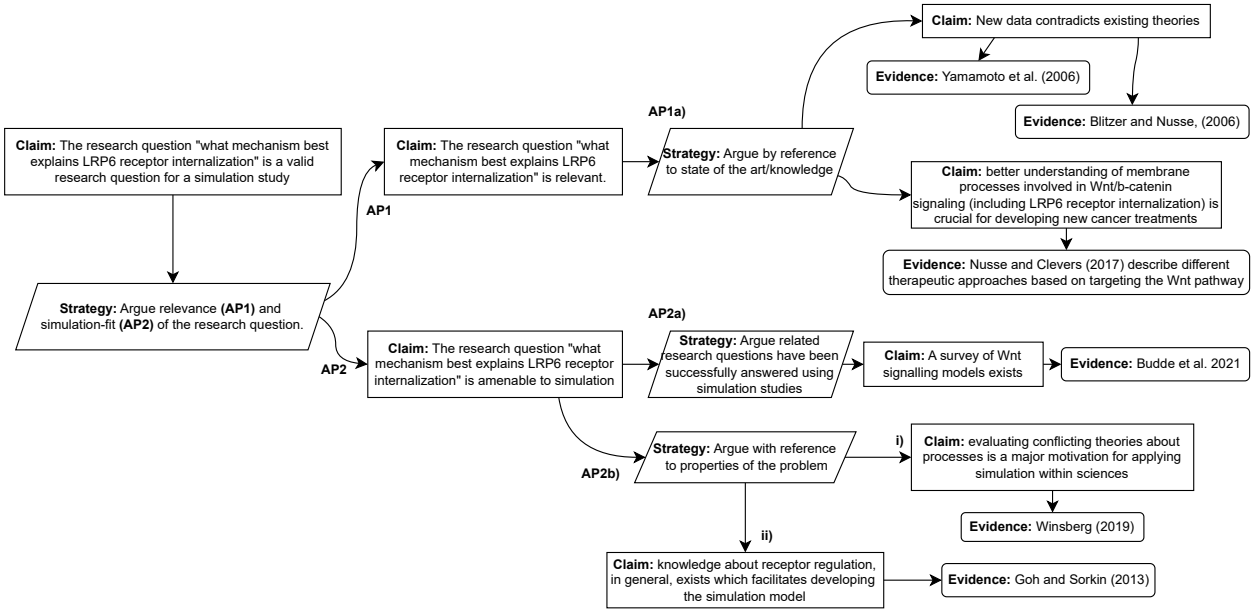

Figure 1: Assurance case for case-study research question. Bold labels indicate parts of the assurance pattern instantiated and link to the description in the text.

canonical signaling? Regarding the **claim** that the research question is relevant, one **argument** based on AP1a) is that new data (Yamamoto et al. 2006) contradicts existing theories reviewed in (Blitzer and Nusse 2006). The references serve as **evidence**. Another **argument** using AP1a) is that a better understanding of membrane processes involved in Wnt/β-catenin signaling (including LRP6 receptor internalization) is crucial for developing new cancer treatments; Nusse and Clevers (2017) discusses different therapeutic approaches based on the Wnt pathway which serves as **evidence** for the claim.

To support the **claim** that *the research question is amenable to simulation*, one **argument** based on AP2a) is that related research questions have been successfully answered using simulation studies, as **evidence** may serve a survey on Wnt signaling models (Budde et al. 2021). We may also **argue** based on AP2b) that i) evaluating conflicting theories about processes is a major motivation for applying simulation within sciences, as **evidence** may serve (Winsberg 2019), and ii) knowledge about receptor regulation, in general, exists which facilitates developing the simulation model, the **evidence** is (Goh and Sorkin 2013).

3.1.2 Requirements

We distinguish between *behavioral (functional) requirements* and *non-functional requirements*, and define specific assurance patterns to support their validity:

Behavioral requirements state what outputs are expected of the model (Ruscheinski et al. 2019). The expectation of a model's output may be given extensionally by data from the modeled physical system or output data from another model that needs to be reproduced by the simulation model. Alternatively, the expected behavior may be specified intentionally – for example, using temporal logic. The behavioral requirements may also be presented as a combination of extension and intention (Piho and Hillston 2021). We propose three assurance patterns:

AP3. **Claim:** *the requirement is relevant to the research question.* We can a) **argue** this claim by stating that the system under study has been observed to behave that way in situations that are of relevance to the research question. **Evidence** can be provided by referencing publications and/or additional provenance about the observed data; or b) **argue** that the modeled system belongs to a class of systems for which prototypical behavior patterns or constraints have been defined, e.g., in terms of

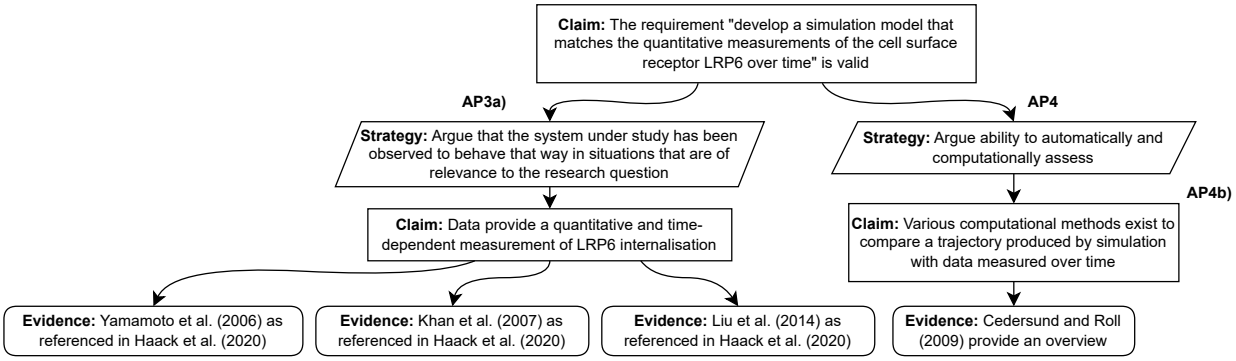

Figure 2: Argument for the validity of simulation requirement R1 "develop a simulation model that matches the quantitative measurements of the cell surface receptor LRP6 over time".

stylized facts (Wilsdorf et al. 2023). **Evidence** can be given by referencing specified behavioral patterns, or constraints.

AP4. **Claim:** *the requirement can be assessed (ideally computationally and automatically).* We may a) **argue** that face validation is an option if we have **evidence** of the availability of multiple experts; or b) otherwise we should be able to **argue** that the requirement's satisfaction can be tested through analysis or simulation experiments. Again, literature will provide the required **evidence** for this claim and argument, possibly (depending on the type of requirement) complemented by a link to an actual simulation experiment artifact.

AP5. **Claim:** *the set of requirements is suitable.* A simulation model fulfilling all requirements should form a suitable basis to answer the research question. **Arguments** need to refer to the system under study and the research question and may elaborate on the requirements' specificity, coverage of the problem, topicality, and coherence. Depending on the type of argument, the provenance of the requirements needs to be specifically evaluated as **evidence**.

The behavioral requirement of the Wnt endocytosis **case study** (see Fig. 2) is to develop a simulation model that matches the quantitative measurements of the cell surface receptor LRP6 over time. Following AP3a), we **argue** the **claim** that it is *relevant to the research question* as the data provide a quantitative and time-dependent measurement for LRP6 internalization. The **evidence** is given by three independent publications, all showing the same behavior; see references given in (Haack et al. 2020). Regarding the second **claim**, we **argue** based on AP4b) that various computational methods exist to compare a trajectory produced by simulation with data measured over time, as shown in Cedersund and Roll (2009). Since the study only defines one requirement, AP5 does not apply.

Non-functional requirements constrain, for example, the runtime of the model or its ease-of-use and visualization (Robinson 2008b). In addition, general project constraints referring to the modeling languages and tools used may be defined, as well as the time frame of the modeling project. We propose an assurance pattern to validate non-functional requirements:

AP6. **Claim:** *the requirement refers to the needs of the project.* Here, we can again **argue** referring to Balci (2012)'s four P's: a) the people involved (**evidenced** by interviews); b) processes involved (literature); c) the product to be built (characteristics of the domain model); or d) the project's characteristics (such as study type and criticality). The precise **arguments**, however, may vary and so will the respective **evidence**

A non-functional requirement of the **case study** has been to develop the model in ML-Rules. This requirement accommodates the needs of the project (**claim**). Following AP6, we can **argue** by looking at the four P's and the associated substrategies a–c): a) the modeler had experiences with the tool from a prior study, b) diverse simulation experiments are supported by the binding (Warnke et al. 2018) between SESSL and ML-Rules, and c) the simulation model to be built needs to include stochasticity and compartmental

dynamics – both supported by ML-Rules (Helms et al. 2017). The respective literature references give **evidence** for these arguments.

3.1.3 Inputs and Outputs

Before specifying the scope and content of the model, it is crucial to identify and define the inputs and outputs (Robinson 2008b). Inputs are the experimental factors through which the modeled system can be exerted to generate outputs, and which can be varied during a simulation experiment. Inputs may be based on the physical system of interest but are also used for hypothetical "what-if" scenarios. They may be quantitative (e.g., rate constants or initial concentrations) or qualitative (i.e., model structures to be varied). It is useful to predefine ranges and distributions of values or possible structural changes. Inputs may also be provided by loading external data, e.g., time series input as a spreadsheet. Output, on the other hand, makes the variables observed during the simulation and possibly subject to behavioral requirements (see above) explicit. We propose two assurance patterns:

AP7. **Claim:** *inputs and outputs are fit to problem.* An important aspect of validating the inputs and outputs artifact is whether inputs and outputs are suitably selected for the system of interest and to answer the research question, taking the assumptions and the model content (see below) into account. Thus, **arguments** and **evidence** might refer to other artifacts within the simulation study besides external sources. To underline the suitability of input values or ranges, publications or provenance might provide evidence, and for outputs, we might wish to inspect the relation to the research question (and behavioral requirements).

AP8. **Claim:** *inputs are reliable.* Here, we can, for example, a) **argue** about the provenance of the data or information: has it been published in a reputable venue? **Evidence** supporting this might include quality indices or certificates; or b) **argue** about data acquisition: Has there been a documented, well-structured, and replicable data-collection process? Has it been replicated independently? Supporting **evidence** could include written reports, lab protocols, or formal provenance graphs. This pattern does typically not apply for hypothetical "what-if" scenarios.

The **case study** defines the rate constants $ke_{nonraft}$ and ke_{raft} as input with parameter ranges of $0.05-0.4\text{min}^{-1}$ and $0.05-0.1\text{min}^{-1}$, respectively. These inputs can be deemed fit to address the research question (**claim**): we employ AP7 to **argue** that they determine the speed of clathrin-independent (nonraft) and clathrin-mediated (raft) internalization of LRP6 receptors – the central processes to be investigated. The survey by Goh and Sorkin (2013) can be referenced as **evidence**. The **claim** concerning the reliability of the inputs could be **argued** using AP8a) by referring to the number of diverse experimental studies cited in the survey (**evidence**). In addition, to **argue** according to AP8b), we could check whether the experimental setup fits the research question addressed in the study (e.g., which cell line has been used for the experiments, and whether the insights from in-vitro experiments were confirmed in-vivo). Provenance information of the studies cited in (Goh and Sorkin 2013) serves as **evidence**, in particular, supplementary materials belonging to the papers may be of value.

3.1.4 Assumptions and Simplifications

In determining the scope and level of detail, various assumptions and simplifications are typically made (Robinson 2008b). Assumptions are made when there are uncertainties or beliefs about the real world. In contrast, simplifications are choices to model something more simply while maintaining a certain level of accuracy. Assumptions and simplifications may involve setting variables to be constant during simulation, eliminating variables, assuming linear relations, removing entire components, or replacing model parts with random variables (Robinson 2008a). They will impact the model's behavior and should not threaten the use of the simulation results to answer the research question. We propose this assurance pattern:

AP9. **Claim:** *the assumption/simplification is sensible* within a simulation study considering the system under study and the research question. Several **arguments** can be made in support of this claim:

a) people: we can **argue** that we have identified the assumptions/simplifications together with experts in the domain and experts in the chosen simulation modeling approach. **Evidence** may include interview scripts or technical reports; b) process: we can **argue** that we have undertaken separate experiments to estimate the impact of simplifications on overall simulation results – for example, through sensitivity analysis, or by experimenting with different spatial resolutions which, then, provide the **evidence**; or c) prior knowledge: we can **argue** that certain assumptions/simplifications have been previously validated in different studies or are commonly considered valid in the domain. **Evidence** can be provided by scientific literature or experiment documentation.

In the **case study**, Haack et al. (2020) assume that "approximately 30% of the membrane is occupied by membrane microdomains (lipid rafts) [...] and, because LRP6 is homogeneously distributed in the membrane, on average 30% of total LPR6 receptors are located in lipid rafts". We apply AP9c) for arguing the **claim** that it is a sensible assumption. Given the research question, i.e., which mechanisms CME or CIE mediates the internalization of LRP6, we **argue** that localization in continuous space or spatial clustering appears of little relevance. Hence, a more fine-grained spatial consideration beyond compartments does not appear meaningful. This is also reflected in the central data sources (Yamamoto et al. 2006) (**evidence**). As a simplification, Haack et al. (2020) states that the interaction between LRP6 and frizzled (FZ) is modeled implicitly. Again, we can **argue** the **claim** according to AP9c). Literature suggests that the Wnt ligand binds to the receptors FZ and LRP6. The resulting trimeric complex recruits Dishevelled (DVL) and Axin, which results in inhibition of β-catenin phosphorylation and β-catenin degradation. FZ is typically abundantly available and thus can be ignored, and the binding can be simplified to LRP6. **Evidence** can be found in Bourhis et al. (2010).

3.1.5 Model Content

The content of a simulation model, also named the domain model (Stepney and Polack 2018), can be subdivided into several components. To identify the content of a simulation model, we need to identify the model boundary and all components within that boundary that we wish to consider and at which level of abstraction. To specify the content of the model, various approaches have been proposed, ranging from formal representations using Petri nets and semi-formal UML class diagrams to informal sketches of the components and their interactions (Wilsdorf et al. 2020).

Validating the model contents is really about validating that we have made the right choices about what to include or exclude from the model – the reduction that is a key property of any model (Stachowiak 1973). This cannot be validated without running simulation experiments requiring a simulation model. We, therefore, adjourn the validation of the model contents – and the appropriate assurance patterns – until we have a simulation model, which will be discussed in the next subsection.

3.2 Simulation Models

"A model (M) for a system (S) and an experiment (E) is anything to which E can be applied to answer a question about S" (Cellier 1991). Thus, a simulation model is designed not only to approximate a system S but also to conduct simulation experiments with a specific research question in mind. More precisely, the simulation model is considered to implement the conceptual model in an executable (and ideally formally defined) modeling language. We propose two assurance patterns for simulation models:

AP10. **Claim:** *the implementation is verified.* Here we need to **argue** that the simulation model (plus the corresponding simulation engine) correctly implements the content, considers the assumptions, etc., as specified in the conceptual model. Arguments about how the content and assumptions have been mapped into the simulation model may refer to the discretization, code structure, use of simulation tools/libraries/frameworks, numerical approximation algorithm, etc., or whether and how parts of the simulation model were generated automatically from the conceptual model (e.g., using

model-driven engineering (Zschaler and Polack 2023)). **Evidence** should outline the respective model transformation process, whether manual or automated.

AP11. **Claim:** *the model is valid.* Here we need to substantiate that the right model has been built. The literature referring to how to validate a simulation model is vast (Balci 1997). In this context, we would like to **argue** that the behavioral requirements are met, by executing suitable simulation experiments (Ruscheinski et al. 2019). **Evidence** is then given by reference to the experiment specification and the simulation data. Successful execution of the experiments also immediately supports the validation of the model content (domain model), if the claim of AP10 applies.

In the **case study**, regarding the **claim** that the model content (the domain model) has been correctly mapped into the simulation model, one could **argue** based on AP10, that due to the rule-based formalism (ML-Rules) used in the implementation, the mapping was straightforward and so is the **evidence**. Reactions as depicted in the content (as a sketch of variables and arrows between them, annotated with kinetics rates), could be directly implemented in ML-Rules. Referring to the **claim** of validating the simulation model, a number of model configurations were scanned. The model configuration that eventually fulfilled the behavioral requirement, i.e. reproduced the experimental in-vitro results, was **argued** to be valid with respect to the conceptual model, following AP11. **Evidence** is provided by the simulation experiments, whose assurance cases will be described in the next subsection. An additional assurance case can be made to **claim** validity as well as more general applicability and predictive power of the model: again using AP11, we **argue** that the simulation model was also cross-validated, i.e., experimental measurements from a different in-vitro experiment that were not part of the calibration and model development process were reproduced by the simulation model. Together with the reference to the independent data set, the corresponding simulation experiment may serve as **evidence**.

3.3 Simulation Experiments

A simulation experiment is the process of generating data from a simulation model by exerting its inputs (Cellier 1991). Simulation experiments are crucial for calibration, validation, and analysis of the simulation model (Budde et al. 2021). The choice and design of experiment depend on the conceptual model, particularly, the research question and requirements defined. Common experiment types include parameter scans, simulation-based optimization, statistical model checking, and sensitivity analysis. To argue the validity of simulation experiments, we propose three assurance patterns:

AP12. **Claim:** *the experiment design is suitable.* We need to make various arguments, e.g., a) we need to **argue** that we chose the right type of experiment for the purpose of testing requirement X, e.g., by referring to prior knowledge as **evidence**; and b) we have to provide **arguments** about the validity of the specific methods employed within the context of this experiment, e.g., methodological literature can serve as **evidence** for confirming whether the chosen sampling approach is appropriate for the number of input factors and assumed dynamics of the model (Pianosi et al. 2016).

AP13. **Claim:** *the experiment is executed correctly.* We need to **argue** that the experiments have been executed correctly. One of the most typical arguments in this context is about aleatory uncertainty: aleatory analysis (Alden et al. 2013) allows determining the appropriate number of executions of a stochastic simulation experiment to ensure any purely random differences in results are averaged out in the results. **Evidence** should refer to methodological papers or preliminary simulation experiments with the hyperparameters, and possibly assumptions or non-functional requirements.

As **evidence** for valid simulation experiments, the **case study** may refer to its provenance graph (given in (Haack et al. 2020)). It provides all relevant information about the simulation experiments that have been executed in the course of the simulation study including their specifications. Furthermore, it refers to other artifacts, such as assumptions, data, input/output specifications, and their justification by references to prior or related work and literature. More specifically, provenance allows us to build assurance cases based on the assurance patterns above. To support the **claim** that the experiment design is suitable, we need to provide two **arguments**. With AP12a), we **argue** that parameter scans are a suitable choice of

experiment type for checking the requirement defined above. **Evidence** is linked with the assurance case of the behavioral requirement and refers to the same methodological literature (Cedersund and Roll 2009). With AP12b), we **argue** that a full factorial scan over the parameters $ke_{nonraft}$ and ke_{raft} is suitable and that the experimental factors are varied within parameter ranges justified by experimental measurements obtained from literature. **Evidence** is linked to the input/output definition as well as the methodology by (Cedersund and Roll 2009). The second **claim** refers to the correct execution of the experiment. Using AP13, we can **argue**, for instance, that the simulation replications were set to a sufficiently large number to provide a confidence level of 95%. Preliminary experiments serve as **evidence**.

4 DISCUSSION

We have presented an analysis of the different artifacts involved in a simulation study and the related needs to assure validity for each artifact. In doing so, we have explored the suitability of using structured arguments to establish and document the validity of different types of artifacts and, as a result, the validity of a simulation study as a whole. We have seen how different artifacts, including those of the conceptual model, require different types of arguments. Overall, structured arguments – as promoted in the field of safety-critical systems – can be applied to document the validity of all the different types of artifacts produced as part of a simulation study. Indeed, the flexibility to choose different argument strategies underpinned by evidence ranging from the more informal to the highly formal appears to be a particular strength of this approach in documenting the validity of a simulation study. With this, they may provide additional structures for documentation guidelines of simulation studies such as TRACE (incl. ODD for model description) (Grimm et al. 2014) and STRESS (Monks et al. 2019). Applying assurance patterns to all artifacts of the simulation study, similar to the V&V triangle of (Brade 2000) and to (OUSD(R&E) 2011), augments these approaches, which focus validation primarily on the simulation model.

With all other documentation standards, structured arguments share the challenge of the induced significant addition effort put on the modeler (Uhrmacher et al. 2024). Another limitation of structured arguments is that they are fundamentally static and hierarchical (at least in the form typically used in safety assurance cases). While this structure is appropriate for arguing the validity of simulation artifacts in isolation, it makes arguing the validity of an overall simulation study more difficult:

1. *Hierarchy and cross-linking:* Arguing overall study validity requires linking together validity arguments for all the artifacts in a study. Figure 3 illustrates this cross-referencing. The hierarchical form of the arguments used in safety assurance makes such cross-linking difficult, typically leaving only informal references in the text of individual claims or evidence.
2. *Static arguments and study evolution:* As a simulation study evolves, we possibly refine the inputs, requirements, assumptions, and of course, the simulation model, as has been done also in (Haack et al. 2020). The static structure of validity arguments means that this process perspective can easily get lost, and thus only part of the story of a simulation study is being told (i.e., the finally approved artifacts with their structured arguments). GSN v3 includes a "dialectic extension" (Assurance Case Working Group (ACWG) 2021, *pp.* 50 *ff.*) to express claims that are countered by other claims, but this still only allows capturing static arguments.

The first challenge can be overcome by moving to a more general, graph-based notion of argument frameworks (building on Toulmin's (1958) original work and work on argument frameworks (Dung 1995)). The second challenge is more complex. Provenance-based approaches (e.g., provenance graphs (Budde et al. 2021)) capture the process perspective, documenting how models were derived as refinements of other models, for example. However, they miss the rationale of what led to these refinements and how they affected the evolution of the diverse artifacts of the simulation study, including research questions, requirements, and simulation models. In mathematics, Pease et al. (2017) have explored the use of agent-dialogue games (Atkinson et al. 2006) as a means of capturing both the process of how a mathematical proof develops as well as the final proof as an argument framework. It would be interesting to explore whether and how this approach could also be applied to documenting the validity of simulation studies as they evolve over

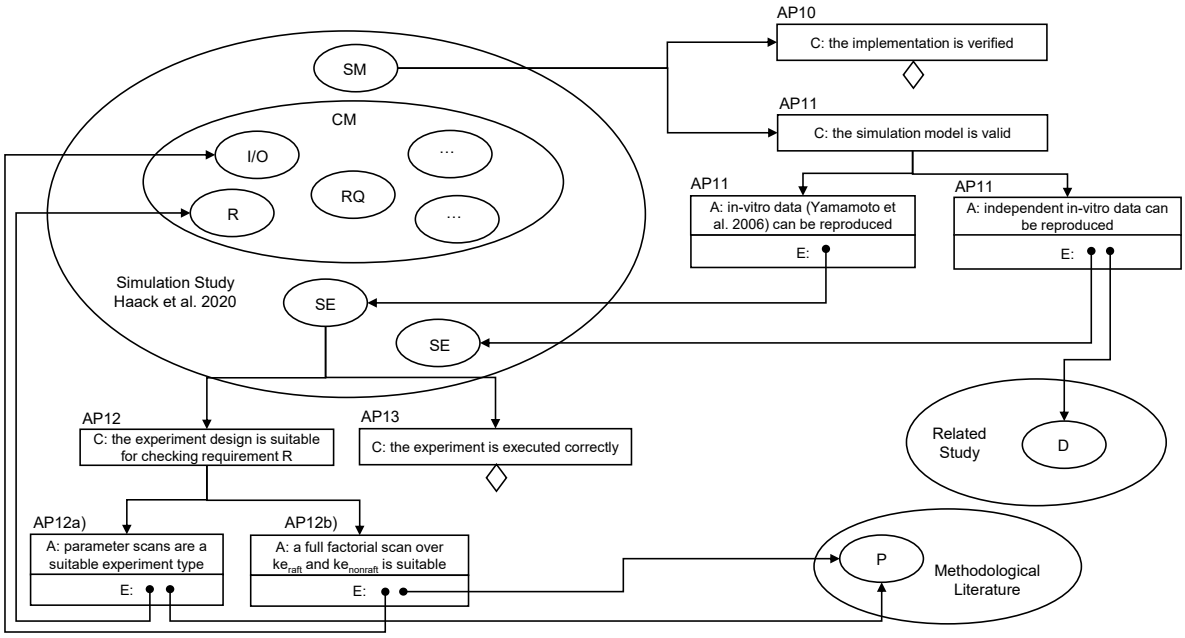

Figure 3: Interdependencies between the artifacts of the case study and their corresponding assurance cases. Each artifact (RQ – research question, R – requirement, I/O – inputs and outputs, SM – simulation model, SE – simulation experiment) is validated using assurance cases (C – claim, A – argument, E – evidence), where evidence may reference one or many other artifacts that live either within the same (e.g., an SE) or a related simulation study (e.g., D – data) or are part of the methodological literature (P – publication).

time. This would require identifying a set of agent-dialogue moves and rules for their application that can be used to model the development of simulation studies and their resulting scientific arguments.

ACKNOWLEDGEMENTS

A.U. acknowledges funding by the German Research Foundation (DFG) grant 320435134 “GrEASE”. F.H. is funded by the German Research Foundation (DFG) grant SFB 1270/2 – 299150580 “ELAINE”.

REFERENCES

Alden, K. 2012. *Simulation and Statistical Techniques to Explore Lymphoid Tissue Organogenesis*. Ph. D. thesis, University of York.

Alden, K. *et al*. 2013. “Spartan: A Comprehensive Tool for Understanding Uncertainty in Simulations of Biological Systems”. *PLOS Computational Biology* 9(2):1–9.

Assurance Case Working Group (ACWG) 2021, May. “Goal Structuring Notation Community Standard Version 3”. Online: https://scsc.uk/SCSC-141C, last accessed 28 February, 2024.

Atkinson, K., T. Bench-Capon, and P. McBurney. 2006, September. “Computational Representation of Practical Argument”. *Synthese* 152(2):157–206.

Balci, O. 1997. “Verification Validation and Accreditation of Simulation Models”. In *Proceedings of the 29th Conference on Winter Simulation*, 135–141 https://doi.org/10.1145/268437.268462.

Balci, O. 2012. “A Life Cycle for Modeling and Simulation”. *SIMULATION* 88(7):870–883.

Blitzer, J. T. and R. Nusse. 2006. “A Critical Role for Endocytosis in Wnt Signaling”. *BMC Cell Biology* 10:1–10.

Bourhis, E., C. Tam, Y. Franke, J. F. Bazan, J. Ernst, J. Hwang, , *et al*. 2010. "Reconstitution of a Frizzled8·Wnt3a·LRP6 Signaling Complex Reveals Multiple Wnt and Dkk1 Binding Sites on LRP6". 285(12):9172–9179.

Brade, D. 2000. "Enhancing Modeling and Simulation Accreditation by Structuring Verification and Validation Results". In *Proc. 2000 Winter Simulation Conference*, 840–848 https://doi.org/10.1109/WSC.2000.899882.

Budde, K., J. Smith, P. Wilsdorf, F. Haack and A. M. Uhrmacher. 2021. "Relating Simulation Studies by Provenance—Developing a Family of Wnt Signaling Models". *PLOS Computational Biology* 17(8):e1009227. Publisher: Public Library of Science.

Cedersund, G. and J. Roll. 2009. "Systems Biology: Model based Evaluation and Comparison of Potential Explanations for given Biological Data". *The FEBS Journal* 276(4):903–922.

Cellier, F. E. 1991. *Continuous System Modeling*. Springer.

Dung, P. M. 1995. "On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and *n*-Person Games". *Artificial Intelligence* 77(2):321–357.

Fujimoto, R., C. Bock, W. Chen, E. Page and J. H. Panchal. 2017. *Research Challenges in Modeling and Simulation for Engineering Complex Systems*. Springer.

Goh, L. K. and A. Sorkin. 2013. "Endocytosis of Receptor Tyrosine Kinases". *Cold Spring Harbor Perspectives in Biology* 5(5):a017459.

Goldschmidt, G. and B. Matthews. 2022. "Formulating Design Research Questions: A Framework". *Design Studies* 78:101062.

Grimm, V., J. Augusiak, A. Focks, B. M. Frank, F. Gabsi, A. S. Johnston *et al*. 2014. "Towards better Modelling and Decision Support: Documenting Model Development, Testing, and Analysis using TRACE". *Ecological Modelling* 280:129–139.

Haack, F., K. Budde, and A. M. Uhrmacher. 2020, August. "Exploring Mechanistic and Temporal Regulation of LRP6 Endocytosis in Canonical WNT Signaling". *Journal of Cell Science* 133(15).

Helms, T., T. Warnke, C. Maus, and A. M. Uhrmacher. 2017, may. "Semantics and Efficient Simulation Algorithms of an Expressive Multilevel Modeling Language". *ACM Trans. Model. Comput. Simul.* 27(2).

Kelly, T. P. and J. A. McDermid. 1997. "Safety Case Construction and Reuse Using Patterns". In *Safe Comp 97*, edited by P. Daniel, 55–69: Springer London.

Knight, J. C. 2002. "Safety Critical Systems: Challenges and Directions". In *Proc. 24th Int'l. Conf. on Software Engineering (ICSE'02)*, 547–550.

Monks, T., C. S. Currie, B. S. Onggo, S. Robinson, M. Kunc and S. J. Taylor. 2019. "Strengthening the Reporting of Empirical Simulation Studies: Introducing the STRESS Guidelines". *Journal of Simulation* 13(1):55–67.

Nusse, R. and H. Clevers. 2017. "Wnt/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities". *Cell* 169(6):985–999.

Object Management Group (OMG) 2021, June. "Structured Assurance Case Metamodel v2.2".

OUSD(R&E) 2011. "Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG)". Online: https://www.cto.mil/sea/vva_rpg/, last viewed 7 June 2024.

Pease, A., J. Lawrence, K. Budzynska, J. Corneli and C. Reed. 2017. "Lakatos-style Collaborative Mathematics through Dialectical, Structured and Abstract Argumentation". *Artificial Intelligence* 246:181–219.

Pianosi, F., K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson *et al*. 2016. "Sensitivity Analysis of Environmental Models: A Systematic Review with Practical Workflow". *Environmental Modelling & Software* 79:214–232.

Piho, P. and J. Hillston. 2021. "Combining Quantitative Data with Logic-based Specifications for Parameter Inference". In *International Symposium: From Data to Models and Back*, 121–137. Springer.

Robinson, S. 2008a. "Conceptual Modelling for Simulation Part I: Definition and Requirements". *Journal of the Operational Research Society* 59(3):278–290.

Robinson, S. 2008b. "Conceptual Modelling for Simulation Part II: a Framework for Conceptual Modelling". *Journal of the Operational Research Society* 59(3):291–304.

Ruscheinski, A., T. Warnke, and A. M. Uhrmacher. 2019. "Artifact-based Workflows for Supporting Simulation Studies". *IEEE Transactions on Knowledge and Data Engineering* 32(6):1064–1078.

Ruscheinski, A., P. Wilsdorf, M. Dombrowsky, and A. M. Uhrmacher. 2019. "Capturing and Reporting Provenance Information of Simulation Studies based on an Artifact-based Workflow Approach". In *Proceedings of the 2019 ACM SIGSIM Conference on Principles of Advanced Discrete Dimulation*, 185–196.

Stachowiak, H. 1973. *Allgemeine Modelltheorie*. Springer.

Stepney, S. and F. A. C. Polack. 2018. *Engineering Simulations as Scientific Instruments: A Pattern Language*. Cham: Springer.

Toulmin, S. 1958. *The Uses of Argument*. Cambridge University Press.

Uhrmacher, A. M., P. Frazier, R. Hahnle, F. Klugl, F. Lorig, B. Ludäscher *et al*. 2024. "Context, Composition, Automation, and Communication - The C2AC Roadmap for Modeling and Simulation". *ACM Transactions on Modeling and Computer Simulation*. Preprint https://doi.org/10.48550/arXiv.2310.05649.

Warnke, T., T. Helms, and A. M. Uhrmacher. 2018. "Reproducible and Flexible Simulation Experiments with ML-Rules and SESSL". *Bioinformatics* 34(8):1424–1427.

Wei, R., T. P. Kelly, X. Dai, S. Zhao and R. Hawkins. 2019. "Model based System Assurance using the Structured Assurance Case Metamodel". *Journal of Systems and Software* 154:211–233.

Wilsdorf, P., F. Haack, and A. M. Uhrmacher. 2020. "Conceptual Models in Simulation Studies: Making it Explicit". In *2020 Winter Simulation Conference (WSC)*, 2353–2364: IEEE https://doi.org/10.1109/WSC48552.2020.9383984.

Wilsdorf, P., M. Zuska, P. Andelfinger, A. M. Uhrmacher and F. Peters. 2023. "Validation Without Data - Formalizing Stylized Facts Of Time Series". In *2023 Winter Simulation Conference (WSC)*, 2674–2685 https://doi.org/10.1109/WSC60868.2023.10408388.

Winsberg, E. 2019. *Science in the Age of Computer Simulation*. University of Chicago Press.

Yamamoto, H., H. Komekado, and A. Kikuchi. 2006, August. "Caveolin Is Necessary for Wnt-3a-Dependent Internalization of LRP6 and Accumulation of β-Catenin". *Developmental Cell* 11(2):213–223. Publisher: Elsevier.

Zschaler, S. and F. A. C. Polack. 2023. "Trustworthy Agent-based Simulation: the Case for Domain-specific Modelling Languages". *Software and Systems Modelling* 22(2):455–470.

AUTHOR BIOGRAPHIES

PIA WILSDORF is a Ph.D. candidate in the Modeling and Simulation group at the University of Rostock. Her e-mail address is pia.wilsdorf@uni-rostock.de.

STEFFEN ZSCHALER is Reader in Software Engineering in the Department of Informatics at King's College London. His e-mail address is szschaler@acm.org.

FIETE HAACK is a Postdoctoral researcher in the Modeling and Simulation group at the University of Rostock. His e-mail address is fiete.haack@uni-rostock.de.

ADELINDE M. UHRMACHER is professor at the Institute for Visual and Analytic Computing, University of Rostock, and head of the Modeling and Simulation group. Her e-mail address is adelinde.uhrmacher@uni-rostock.de.