Credibility assessment of computational models according to ASME V&V40: Application to the Bologna Biomechanical Computed Tomography solution

Computer Methods and Programs in Biomedicine 240 (2023) 107727

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Credibility assessment of computational models according to ASME V&V40: Application to the Bologna Biomechanical Computed Tomography solution

Alessandra Aldieri [a,b,*], Cristina Curreli [b,c], Julia Aleksandra Szyszko [b,c], Antonino Amedeo La Mattina [b,c], Marco Viceconti [b,c]

[a] PolitoBIOMedLab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Italy
[b] Medical Technology Lab, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy
[c] Department of Industrial Engineering, Alma Mater Studiorum - University of Bologna, Bologna, Itlay

ARTICLE INFO

Keywords:
Model credibility
ASME V&V40
VVUQ
In silico trial
Hip fracture
Fracture risk prediction

ABSTRACT

Background and objective: When a computational model aims to be adopted beyond research purposes, *e.g.* to inform a clinical or regulatory decision, trust must be placed in its predictive accuracy. This practically translates into the need to demonstrate its credibility. In fact, prior to its adoption for regulatory purposes, an *in silico* methodology should be proven credible enough for the scope. This has become especially relevant as, although evidence of the safety and efficacy of new medical products or interventions has been traditionally provided to the regulator experimentally, *i.e., in vivo* or *ex vivo,* recently the idea to inform a regulatory decision *in silico* has made its way in the regulatory scenario. While a harmonised technical standard is currently missing in the EU regulatory system, in 2018 the ASME issued V&V40–2018, where a risk-based framework to assess the credibility of a computational model through the performance of predefined credibility activities is provided. The credibility framework is here applied to Bologna Biomechanical Computed Tomography (BBCT) solution, which predicts the absolute risk of fracture at the femur for a subject. BBCT has recently been the object of a qualification advice request to the European Medicine Agency.
Methods: The full implementation of ASME V&V40–2018 framework on BBCT is shown. Starting from BBCT proposed context of use the whole credibility plan is presented and the credibility activities (Verification, Validation, Applicability) described together with the achieved credibility levels.
Results: BBCT risk is judged medium, and the credibility levels achieved considered acceptable. The uncertainties intrinsically present in the material properties assignment affected BBCT predictions to the highest extent.
Conclusions: This work provides the practical application of the ASME V&V40–2018 risk-based credibility assessment framework, which could be applied to demonstrate model credibility in any field and support future regulatory submissions and foster the adoption of In Silico Trials.

1. Introduction

Before it can be sold in a country, every new medical product must be proven safe and effective when employed according to its intended use. To grant marketing authorisation for a new medical product, regulatory agencies should be provided with evidence of safety and efficacy, which historically comes from controlled experiments, either performed *in vivo* or *ex vivo*. Nonetheless, due to its power to reduce, refine and replace *in vivo* experimentation and bench tests, modelling and simulation have recently emerged as valid alternatives in selected cases. In fact, regulatory agencies have started to accept evidence coming from modelling and simulation, *i.e., in silico* [1]. When the use of modelling and simulation in medicine is employed to provide evidence of new medical products safety or efficacy, it is referred to as *In Silico* Trial (IST). From a regulatory perspective, the adoption of ISTs to inform regulatory decisions poses unique challenges: the regulatory target is, in fact, represented by a new medical product that the IST solution is expected to provide evidence about. The first requirement in order that such

* Corresponding author at: Department of Industrial Engineering, Alma Mater Studiorum - University of Bologna, Medical Technology Lab, IRCCS Istituto Ortopedico Rizzoli, Bologna (IT), Via di Barbiano 1/10, 40136 Bologna (IT).
E-mail address: alessandra.aldieri@polito.it (A. Aldieri).

https://doi.org/10.1016/j.cmpb.2023.107727
Received 27 January 2023; Received in revised form 17 July 2023; Accepted 18 July 2023
Available online 26 July 2023

evidence obtained *in silico* is accepted is that the *in silico* methodology adopted is qualified. As a first step, the qualification of an *in silico* methodology requires the assessment of its *credibility*. In 1980, credibility for a computational model was defined as the ability to replicate the modelled reality within a predefined tolerance [2]. Recently, credibility definition was expanded to a broader context, and a model is deemed credible if able to elicit belief or trust in its results also accounting for its risk level [3,4].

Although extensive knowledge and well-established practices apply to the qualification of *in vitro* and *in vivo* methods, an analogously accepted framework for credibility has not yet been established for computer modelling. In principle, credibility for an *in silico* model should be established through Verification & Validation (V&V) activities aimed at proving that the model outputs are sufficiently accurate and reliable for a specific application [5]. However, a rigorous and structured framework is paramount when assessing the credibility of an *in silico* model used in a regulatory submission. The regulatory gap represented by the lack of such a framework has become apparent in recent years. This has raised the attention of the computational medicine research community, together with the efforts of some regulators in the USA and Europe. Already in 2014, Food and Drug Administration (FDA) issued the guidance "Reporting of computational modelling studies in medical device submissions" and, in parallel, committed to the establishment of an ASME standardisation committee which led to the publication, in 2018, of the technical standard ASME V&V40 "Verification and Validation in computational modelling of medical devices" [4]. The ASME V&V40–2018 proposes a risk-based framework of credibility activities to consistently elicit evidence on the credibility of a computational model. ASME V&V40–2018 standard was initially conceived aimed to provide a credibility framework for computational models in medical device submissions. Hence, the regulatory focus was primarily on the specific medical device, developed or proved safe and effective using a computational model. Nevertheless, ASME V&V40–2018 adoption in additional fields and for additional scopes was later endorsed. In fact, in its 2021 guidelines [6], FDA endorsed the adoption of ASME V&V40–2018 for the qualification of computational models, specifically referring to its 'Qualification of Medical Device Development Tools' guidance [7]. As far as the EU regulatory system is concerned, there is currently no available harmonised technical standard providing a framework for the credibility assessment of an IST solution, as the ASME V&V40 does. The new European Medical Device Regulation (EU-MDR) only claims that, where appropriate, "the results of biophysical or modelling research the validity of which has been demonstrated beforehand" could be considered in relation to the device requirements regarding design and manufacture [8].

In 2019, the European Medicine Agency (EMA) issued a Guideline on reporting physiologically based pharmacokinetic (PBPK) modelling and simulation [9]. It resulted from the growing number of regulatory submissions including PBPK models, and the subsequent need to frame the information to be included within a PBPK modelling report, as well as to advise on how to qualify a PBPK platform. Modelling spread so much in the pharmacological field that a Modelling and Simulation Working Party (MSWP) was created to support EMA scientific committees and working parties on modelling and simulation relating to medicines. Aware of the value *in silico* methods are gaining in drug development, and of the lack of international standards for their evaluation, the ASME V&V40–2018 standard has recently been endorsed also in the context of drug development and evaluation [10,11]. This proves the flexibility of the framework offered by ASME V&V40–2018 standard, whose possible application in different fields was recognised.

In this context, the aim of this paper is to provide an example of the practical application of the credibility assessment according to the ASME V&V40 standard of an *in silico* model for the prediction of the femur fracture risk. Recently, this *in silico* methodology has been the object of a qualification advice request to EMA concerning its qualification to be adopted to test the efficacy of new treatments against osteoporosis in Phase II clinical studies. In the "Guideline on the Evaluation of Medicinal Products in the Treatment of Primary Osteoporosis" [12], EMA recommends testing new treatments efficacy by adopting the fracture incidence as primary outcome in Phase III clinical trials. However, in Phase II or supportive Phase III studies the use of a surrogate of the fracture endpoint, namely the areal Bone Mineral Density (aBMD) from Dual X-ray Absorptiometry (DXA) is admitted and often adopted as primary outcome. Nevertheless, aBMD does not represent a good fracture predictor, able to explain only 60% to 75% of the variance in the fracture risk [13–15]. The *in silico* methodology called the Bologna Biomechanical Computed Tomography at the hip, BBCT-hip, predicts an absolute risk of fracture which has proved to be a surrogate of the fracture endpoint more accurate than aBMD [17,16]. In a retrospective cohort study where BBCT-hip accuracy in separating fracture from non-fracture cases was compared to aBMD [16], the former showed an area under curve improvement of 12% with respect to the letter (0.87 against 0.75). The first step within a qualification advice request for an *in silico* methodology is to provide evidence of its credibility. In accordance with [10], we therefore evaluated BBCT-hip credibility through the verification and validation activities framed by the ASME V&V40–2018 standard. Starting from the definition of a so-called Context of Use (CoU), an overall risk is determined for BBCT-hip and the resulting verification, validation, uncertainty quantification, and applicability analysis plan is carried out.

2. Materials and methods

The overall credibility assessment workflow that the ASME V&V40–2018 standard recommends is provided in Fig. 1. The core of the pipeline lies in the definition of one or multiple Context(s) of Use (CoU) for the model: it is based on the CoU that the model risk is determined, and consequently the whole credibility plan established.

The specific application of the ASMEV&V40–2018 framework to BBCT-hip case will be detailed in the following sections.

2.1. BBCT-hip model in brief

BBCT-hip model calculates the hip fracture risk upon falling, named ARF0, by modelling a fall to the side. In principle, the fracture risk is identified by calculating possible impact forces derived from a fall (through a stochastic mathematical model) and by assessing which of those, exceeding the load to failure (determined through a patient-specific finite element model), lead to a fracture event [17]. More in detail, BBCT-hip uses a stochastic mathematical model to simulate 1000,000 falls of a body of the height and weight equal to those of the patient, each with initial conditions assigned randomly according to specific probability distributions, and for each of these falls predicts the resulting impact force. While height and weight of each subject represent deterministic inputs of the model, the other inputs required, *i.e.*, initial and final velocity and acceleration, postural and impact attenuation variables, are described by normal distributions truncated symmetrically at $\pm$ 3 standard deviations with respect to the mean. The truncation points for the distributions were taken from the literature [17], as better specified in the Supplementary Material. The 1000,000 values for the impact force due to a fall are drawn by sampling the distributions of the stochastic variables, assumed to be independent, using inverse-transformed Latin Hypercube.

In parallel, a patient-specific Finite Element (FE) model of the femur informed by the patient's Quantitative Computed Tomography (QCT) data is run 28 times, varying the femur orientation at the impact (femoral impact pose). Heterogeneous Hounsfield Units-based material properties are assigned according to [18]. In order to replicate a sideways fall, a concentrated force is applied at the centre of the femoral head; a contact interaction is defined between the greater trochanter surface and a rigid static plane orientated normally to the direction of force; the distal part of the femur (25% of the biomechanical length from

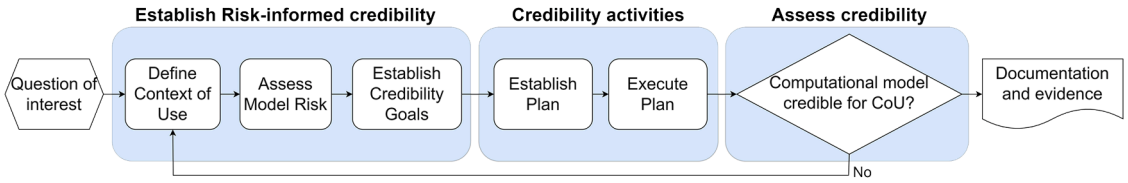

Fig. 1. ASME V&V40–2018 workflow to assess an in silico model credibility. Reprinted from ASME V&V 40–2018 by permission of The American Society of Mechanical Engineers. All rights reserved.

the knee centre) is constrained to a hinge located at the centre of the knee. For each impact pose, the load to failure, *i.e.* the intensity of the force required to fracture the femur, is computed based on principal strains [19] in a region of interest located proximally. The FE model-derived loads to failure inform a reduced-order model (response surface) which allows inferring the magnitude of the load to failure for each possible impact direction at a reasonable computational cost. By comparing the FE-derived loads to failure to the computed impact forces, the absolute risk of hip fracture at time 0 (ARF0), where time 0 refers to the time of the QCT scan, can be computed. More specifically, the surrogate biomarker ARF0 is calculated as the ratio of the number of simulated falls that the model predicts would cause a fracture divided by the total number of simulated falls. Fig. 2 provides a graphical overview of BBCT-hip pipeline.

BBCT-hip is described in greater detail in the Supplementary Material.

2.2. From the Quantity of Interest (QoI)I to the model risk

According to ASME V&V40–2018, the credibility assessment procedure starts with the definition of a (or multiple) scientific Question of Interest (QoI) which will be addressed using the model. In the case of BBCT-hip, the identified QoI was (Fig. 3):

which is the optimal effective dose for a new anti-osteoporosis drug in adults and older adults (from 55 years) according to multi-dose Phase II studies? (Fig. 2).

Starting from how BBCT-hip would be used to 1answer the QoI and inform a regulatory decision, the CoU was defined. The CoU should concisely describe the role and scope of the model in answering the QoI, briefly describe the data used to build the model and the way the model outcomes will be used. The following CoU was proposed for BBCT-hip:

BBCT-hip is a methodology where a stochastic biophysics model provides an estimate, for a given subject, of the Absolute Risk of proximal femur Fracture upon falling at time zero (ARF0), from their height, weight, and a Quantitative Computed Tomography (QCT) scan of the hip region. This ARF0 is to be used as a response variable in multi-dose Phase II studies in place of the measured DXA-based aBMD. The average change in ARF0 over the period of treatment for all subjects treated with a given dose ($Ave_{\Delta ARF0}$) can be used as response variable, by assuming the optimal dose amongst those tested is the one for which $Ave_{\Delta ARF0}$ is most positive (or least negative).

Once the CoU is defined, the model's overall risk can be established. That is crucial since the risk level identified for the model will strongly affect the following necessary credibility activities and the overall effort to demonstrate the model is credible enough for its CoU. In ASME V&V40 the definition of the model risk merges the concepts of decision consequence, *i.e.* the severity of the adverse outcome if the decision based on the model is incorrect, and of model influence, *i.e.* the contribution of the model to the final decision. In order to assess BBCT-hip risk, the decision consequence was considered consistently with the ASME V&V40 definition. On the contrary, the model influence was substituted by the so-called regulatory impact. This pragmatical approach relates to the endorsement and application of ASME V&V 40 beyond its original scope. Indeed, it was suggested in [20], where in the pharmacological context the regulatory impact was introduced aiming to focus the attention on the regulatory decision and more explicitly compare and contrast the model with alternative established methods to answer the same question of interest. Therefore, as it emphasises the regulatory context and the availability of additional sources of evidence to reach the regulatory decision, regulatory impact was here preferred over model influence, which has a very broad definition.

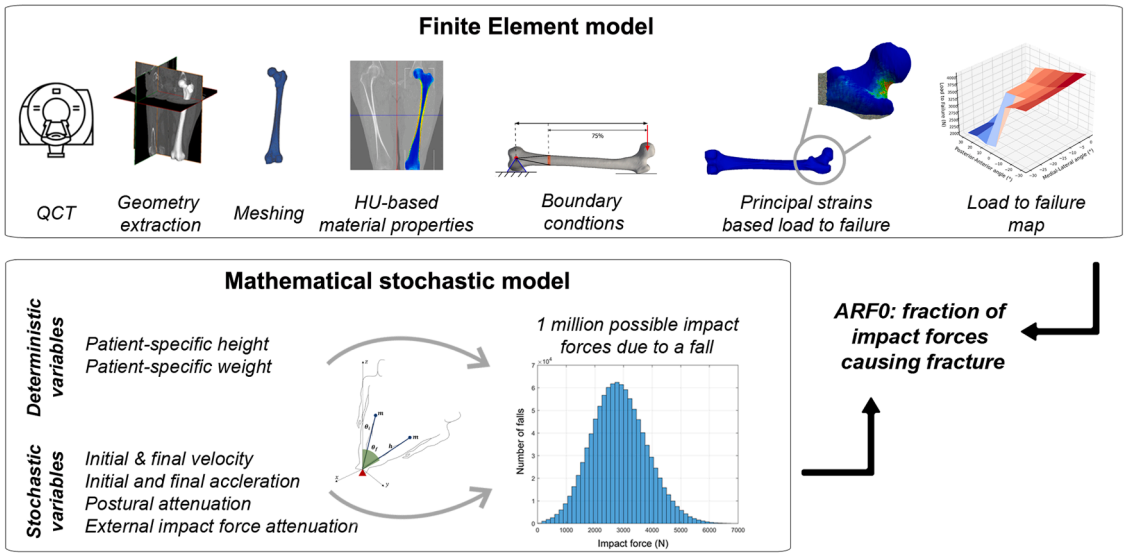

Fig. 2. Graphical overview of BBCT-hip methodology. In the upper panel the FE model construction is presented, which yields the response surface related to the loads to failure. In the lower panel, the input parameters for the stochastic model are listed, which, used within a 1 degree of freedom (DoF) inverse pendulum model, allow to obtain possible impact forces due to a fall on the side. By computing the ratio between the forces which would exceed the loads to failure and thus cause fracture and the total number of forces, ARF0 is calculated.

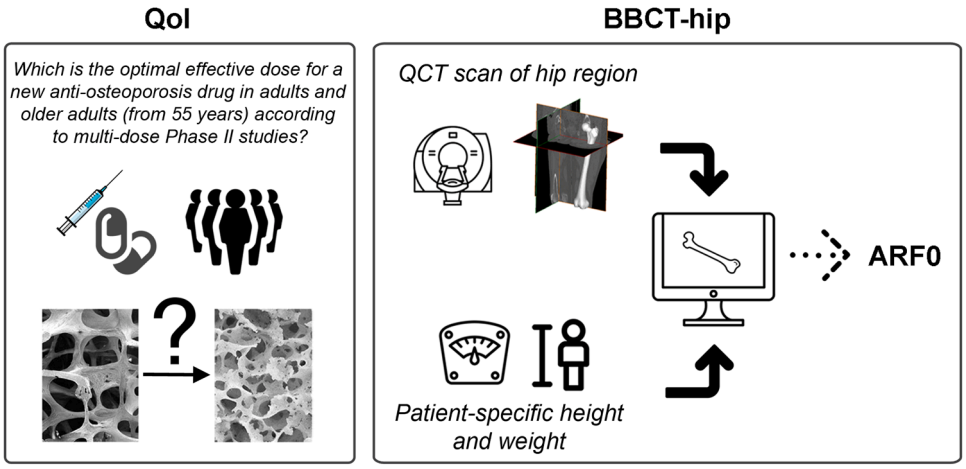

Fig. 3. Left: the QoI that is addressed by BBCT-hip model. Right: a graphical representation of how BBCT-hip methodology operates within the proposed CoU: BBCT-hip provides an absolute risk of hip fracture, which shall be employed as a surrogate biomarker of the fracture incidence to test the efficacy of new treatments against osteoporosis in multi-dose Phase II trials.

For BBCT-hip, decision consequence was considered *low*: an inaccurate decision on the efficacy of a new drug in Phase II might relate to patients being either over- or underexposed to the drug in Phase III studies, but that dose would, in any case, be lower than the maximum tolerable dose and higher than the minimum effective dose, both established during a previous phase I study. Therefore, the patient would not be exposed to any additional risk. Furthermore, the selection of a sub-optimal dose would not prevent an effective drug from making it to the market: if the new molecule is significantly more effective than a comparator the responsiveness registered in a Phase III study would not be affected to this extent. The impact on the regulatory decision was considered *high*: the BBCT-based biomarker (ARF0) is proposed as a substitute of aBMD biomarker as surrogate of the fracture endpoint in multi-dose Phase II clinical studies to evaluate the efficacy of different doses for new antiresorptive treatments. As such, the choice of the optimal dose would be based entirely based on the BBCT-hip prediction. BBCT-hip overall risk was therefore determined to be at level 3 for the proposed CoU, as highlighted in the graphical representation provided by the five-level risk map (Fig. 4). In fact, additional and key evidence for the final regulatory decision would still be available from phase III trials.

2.3. Credibility activities

ASME V&V-40 standard provides a list of so-called credibility activities which should be considered and carried out to assess whether the credibility of the model is sufficient for the context of use and the risk previously defined. Each credibility activity is decomposed in multiple credibility factors, which represent aspects related to the model which should be considered to carry out the V&V activities. Each factor can be taken into account with a rigour (credibility level) which should be commensurate to the model risk.

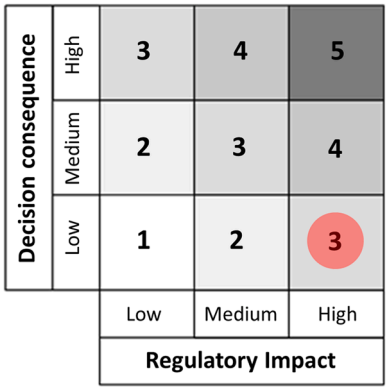

Fig. 4. Five-level risk map where the identified BBCT-hip overall risk has been highlighted as resulting from a high regulatory impact and a low decision consequence.

BBCT-hip estimates ARF0 combining a finite element model, which allows to predict a load to failure based on principal deformations, with a mathematical model that predicts fall impact forces. Hence, the main quantities of interest considered for the V&V activities were: 1) ARF0, 2) the load to failure, 3) the Minimum Side-Fall Strength (MSF), *i.e.,* the lowest load to failure across the 28 femoral impact pose and 3) the deformations on the proximal femur surface. V&V activities were carried out taking advantage of a dataset of 101 calibrated CT scans collected at Rizzoli Orthopaedic Institute (IOR) from 1999 to 2016 for surgical planning of hip arthroplasty at the contralateral femur (with approval from Istituto Ortopedico Rizzoli Ethics Committee), on which BBCT-hip pipeline could be carried out. This will be referred to as the Bologna cohort in the following. Besides, also a retrospective cohort of 98 post-menopausal women, which will be referred to as the Sheffield cohort, was considered to assess the stratification accuracy of BBCT-hip.

2.3.1. Verification: code verification

Code verification (Software Quality Assurance and Numerical Code Verification) aims to quantify the reliability of underlying software implementation, including its numerical accuracy.

Code verification evidence was provided by the vendors of the simulation frameworks used to implement BBCT-hip (Ansys Inc; MathWorks Inc). The verification of Mechanical Ansys Parametric Design Language (APDL) program (Ansys scripting language used to set a simulation and interacting with Ansys Mechanical solver) relied on ANSYS Inc.'s Quality System, developed and evolved to meet ISO 9001 requirements. In addition, two test cases (VM211 and VM184) available in Ansys Mechanical APDL verification manual [21] were used to compare the obtained numerical solution with known analytical solutions. Matlab software verification was based on the availability of detailed audit reports from a third-party independent testing body, TÜV SÜD in Germany. These are provided with the IEC Certification Kit regarding tool certification requirements of IEC 61,508 standard and attest that the software development and validation practices followed by MathWorks adhere to the highest standards in the industry.

2.3.2. Verification: calculation verification

Calculation verifications were conducted on all components of BBCT-

hip which could contribute to the numerical approximation error.

The discretisation error, associated with the computation of the solution at a finite number of points, was estimated with reference to the following features:

- *The Latin Hypercube sample size* (within the stochastic mathematical model) adopted to estimate the impact loads. The sample size was varied from 10^2 to 10^7 and ARF0 computed for each sample size on the whole Bologna cohort.
- *The number of impact poses* simulated with the full-order finite element model to inform the reduced-order model. The intra-extra and add-abduction impact direction angle ranges were sampled in steps of 10°, 5°, and 2° The full-order finite element model run 28, 91 and 496 different times, respectively varying the femoral impact pose for each subject. ARF0 and MSF were computed. The analysis was conducted considering the whole Bologna cohort.
- *Mesh dimension,* considering principal strains. The minimum principal deformation was extracted at the same location for different mesh sizes and the percentage difference was computed for each model with respect to the most refined mesh (0.75 mm size edge size). The FE simulations were run in the neutral impact pose only (0° intra-extra rotation and add-abduction angles). This analysis was run for two different subjects (S1 and S2) belonging to the Bologna cohort. The list of the edge sizes tested is reported in Table S1 in the electronic supplementary material.
- *The bone elasticity spatial discretisation, i.e.,* the number of Hounsfield Unit (HU)-dependent mesh element groups (called material cards in Ansys) used for Young's modulus assignment. The elastic properties of the bone are assumed as heterogeneous over space. This implies a continuum representation, but also, in this case, the finite element solution provides only a spatially discrete description of the bone tissue elasticity, which converges to the exact distribution as the number of individual elasticity values increases. In order to identify an upper boundary estimation of the approximation error due to the discretisation of the otherwise continuous elastic properties, simulations were carried out where a number of material cards able to separate 1 HU difference were implemented, and comparisons were made with the standard material assignment, where a 50 HU difference is considered. The analysis was performed on a single subject (S3) included in the Bologna cohort for each of the 28 impact poses.

The Numerical solver error, *i.e.*, the error that originated from the numerical solutions based on the FE solver parameters was evaluated on the full-order FE model. In particular, a sensitivity study was performed on Newton-Raphson convergence criteria parameters. By default, L2 norm with a 0.005 tolerance is used for convergence check within the Newton-Raphson algorithm for nonlinearity. The effect of the variations of the tolerance and the norm adopted to establish convergence in Newton Raphson method on the load to failure estimated in the Neutral orientation (0° in both intra-extra and add-abduction falling directions) and on ARF0 was quantified. The FE model built starting from one subject (S3) included in the Bologna cohort was considered. The list of the tested tolerances and norms is reported in Table S2 in the electronic supplementary material.

2.3.3. Validation: computational model

Validation activities in this context were all performed on the full order FE model, which, based on principal deformations, predicts the load to failure.

The V&V-40 Standard refers to model form so as to include conceptual and mathematical formulation of the computational model. That involves the form of governing equations, system configuration, system properties and system conditions.

In this case:

- *Governing equations form: the choice of the density to Young's modulus relationship expressions.*

To this end, reference was made to [22], where predictive accuracy on deformation for FE models implementing three distinct power law density-elasticity relationships [18,23,24] was assessed. The three density-elasticity relationships implemented in [24] are listed in the following:

1) $E = 3790 \cdot \rho_{app}^{3}$ [23],
2) $E = 10,500 \cdot \rho_{app}^{2.29}$ [24],
3) $E = 6850 \cdot \rho_{app}^{1.49}$ [18].

where ρ_{app} (in g/cm^3) is the apparent density as derived from the HU of the QCT, and E (in MPa) is the elastic modulus.

- *System configuration form: the CT-derived femur geometry.*

Uncertainties resulting from the procedure followed to isolate the femur geometrical model (*i.e.*, segmentation) were evaluated by quantifying inter- and intra-operator variability on the resulting geometric models. In addition, the effect of such uncertainties on the load to failure and ARF0 was assessed by developing the FE models from the segmentations performed four times by the same operator or from four different operators. A total of 4 different subjects (S1, S2, S3, S4) belonging to the Bologna cohort were considered for this purpose.

- *System conditions form: the applied boundary conditions to simulate a fall on the side.*

Reference was here made to the work of [25], where the implementation of three different boundary conditions for considering the impact at the ground (Linear, Multi-point constraints and Contact model at the greater trochanter) allowed to investigate the effect of different side fall-reproducing boundary conditions on ARF0 and its stratification accuracy on the Sheffield retrospective cohort.

In line with V&V-40 Standard, the sensitivity of the biomechanical quantities of interest to model inputs and the degree to which uncertainties in the model inputs propagated to the model predictions were assessed. In particular, model inputs refer to the values of the parameters used in the governing equations, system configuration, system properties and system conditions. BBCT-hip input parameters included in the analysis are listed in the following.

- *Governing equations inputs: the coefficients of Morgan's relation to determine Young's modulus from density* [18].

A density to elasticity relationship (see Supplementary Material) is adopted to convert the apparent density (ρ_{app} in g/cm^3) as derived from the HU of the QCT images to the elastic modulus (E in MPa) to be assigned to each element of the FE model: $E = a \cdot \rho_{app}^{b}$, where $a = 6850$, $b = 1.49$ are adopted in the above relation. Those values were reported in the work of [18] as mean regression coefficients. In the same work, 95% confidence intervals bounds are reported for coefficients a and b (a: 5440 − 8630; b: 1.14 − 1.84). Those bounds were hence employed to derive material properties from the resulting $\rho_{app} - E$ relations (Fig. 5) and to assess the changes in the quantities of interest (ARF0, loads to failure and MSF). This analysis was carried out considering one only subject (S3) from the Bologna cohort.

- *Governing equations inputs: the coefficients included in the HU-density calibration law to quantify voxel mineral density from their HU.*

QCT images calibration is performed using a phantom, *i.e.*, a body

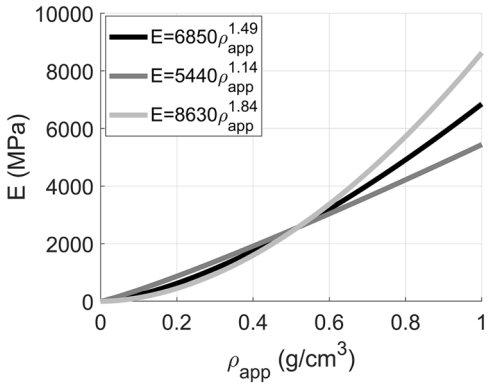

Fig. 5. In black, the $\rho_{app} - E$ relationship provided in [18] and adopted within BBCT-hip. In dark and light grey, the upper and lower bound $\rho_{app} - E$ curves defined based on a and b parameters delimiting the 95% confidence interval.

composed of multiple parts characterised by different calcium hydroxyapatite concentrations (the inserts represent the spongious bone, cortical structure and spinal process) and therefore by different known density values ρ_{QCT}. The phantom is scanned, the mean HU of each portion of the phantom is extracted, and knowing the corresponding real density value, the linear calibration relation is defined with the following form: $\rho_{QCT} = c + d \cdot HU$. The uncertainties affecting the calibration line were quantified by progressively considering an increased or decreased number of pixels (*i.e.,* moving 2 mm towards/away from the outer edges of the portion) for averaging the phantom HU.

The uncertainty area and the mean regression line are shown in Fig. 6, while the calibration line coefficients are reported in Table S3 (electronic supplementary material). The effects on the model outcomes (ARF0, loads to failure and MSF) were quantified on one subject (S3) included in the Bologna cohort by employing the different calibration lines to determine density from HU.

- *System configuration inputs: the locations of the anatomical landmarks identified on the femur to create the anatomical reference system used for load and boundary conditions.*

Reference anatomical landmarks identified by the user at the femur epicondyles are used within BBCT-hip to define a local anatomical reference system used in the FE simulations (see Supplementary Material). Hence, intra- and inter-operator uncertainties are assessed: the epicondyle points are detected 1) by the same operator four different times and 2) by four different operators. A different local reference system is defined for each set of landmarks identified, and the effect on the quantities of interest (ARF0, loads to failure and MSF) is quantified on one single subject (S3) from the Bologna cohort.

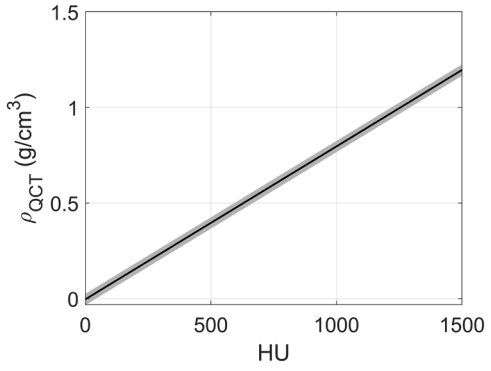

Fig. 6. The calibration line employed to derive density from the CT Hounsfield Units. The shaded area highlights the uncertainty associated to the average over a varying number of HU.

- *System conditions inputs: the portion of the distal diaphysis, computed as a percentage of the total biomechanical length, constrained to rotate around the knee centre.*

In the FE model the femur is distally constrained to rotate around the knee centre (see Supplementary Material). In particular, all the distal nodes positioned below 25% of the full biomechanical length with respect to the knee centre are selected for the constraint application. To assess how the choice of this specific distance impacts on the simulation outcomes, the same constraint was applied selecting the nodes positioned below 5 and 45% respectively of the full biomechanical length. The effects on BBCT-hip results (ARF0, loads to failure and MSF) were assessed considering one subject (S3) from the Bologna cohort.

- *System conditions inputs: the contact parameters employed in the definition of the rigid frictionless contact plane placed at the greater trochanter.*

A rigid contact plane is placed at the greater trochanter in the finite element model to simulate contact with the ground during the impact (see Supplementary Material). A number of real constants are used by Ansys Mechanical APDL in the contact plane definition. The main two, the normal penalty stiffness factor and the penetration tolerance value, are taken into account: starting from the default values of 1 and 0.1 for the normal penalty stiffness factor and the penetration tolerance value, respectively, the effect of their variation (Table S4 in the electronic supplementary material) on the quantities of interest (ARF0, loads to failure and MSF) is quantified on one subject (S3) included in the Bologna cohort.

In addition, the combined effects of the uncertainties in the CT-derived femur geometry, CT calibration parameters and location of anatomical landmarks are considered. The effect of such variations on the on the quantities of interest (ARF0, loads to failure and MSF) is evaluated on one subject (S3) included in the Bologna cohort. The simulations have been run by combining in all possible ways the values as independently set in the previously presented uncertainty analyses.

2.3.4. Validation: comparator – observed data

According to ASME V&V-40 terminology the experimental comparator represents the experimental evidence which has to be compared to the outcome predicted by the model in order to assess its predictive accuracy and complete validation. In accordance to [20], we will here refer to observed data as well, which better fits the assessment of the stratification accuracy in relation to this credibility activity.

In fact, because BBCT-hip is composed of multiple components, validation needs to be carried out on multiple levels. Moreover, the *in silico* methodology BBCT-hip is employed to predict the risk of fracture *in vivo* instead of specific performances of medical devices. This forced validation activities to be performed from a twofold perspective, as detailed in the following.

- *Prediction accuracy*: BBCT-hip predicts the load to failure by first predicting the biomechanical deformation induced in the bone tissue by the loading conditions. Prediction accuracy was therefore assessed by comparing these biomechanical quantities of interest, namely deformations and load to failure, to *ex vivo* experimental studies. Cadaver femurs were scanned and used to inform a QCT-based finite element model. Experimental tests were performed on the same femurs, and analogous boundary conditions were replicated between the experiment and the model. The predictions of the model were compared to the experimental measurements obtained from strain gauges for validation. In addition, by loading the femur until fracture, the maximum applied load which caused the fracture

and the anatomical point where the fracture originated were recorded and compared to the model predictions. Reference will be here made to publications where those comparisons were presented to validate the FE pipeline [26].

- *Stratification accuracy*: aiming to assess whether a good accuracy in the prediction of biomechanical quantities (deformation and load to failure) does translate into a higher accuracy of BBCT-hip in predicting the risk of hip fracture in the clinical practice, ARF0 retrospective stratification accuracy was evaluated. In [17,16], BBCT-hip was employed on the Sheffield cohort, composed of 98 post-menopausal women, 49 of whom having experienced a hip fracture. The control subjects were pair-matched with the fractured ones by age, height, weight, T-score. Average age was 75 years and the average T-score −1.4. ARF0 accuracy in separating fractured from non-fractured patients was assessed and compared to that of the gold standard aBMD.

After the outputs of the V&V activities are obtained and compared the robustness of those activities is evaluated (**Assessment**). That is done considering type, number and equivalency of the compared variables as well as the rigour and degree of agreement of their comparison.

2.3.5. Applicability

The credibility of a model can be demonstrated as long as the validation activities and the evaluated quantities of interest are relevant to its use according to the predetermined CoU. This is the reason why Applicability is part of ASME V&V40–2018 framework. In the context of BBCT-hip, specific reference was made to how its predictive and stratification accuracy was determined.

So, the *Relevance of the quantities of interest* considered in the validation studies and the *Relevance of the validation activities to the CoU* were evaluated.

4. Results

Table 1 presents the credibility activities and the corresponding credibility factors which were considered. The rigour selected for each factor, and the resulting credibility level achieved, are also reported in accordance with ASME V&V 40–2018. ASME V&V 40–2018 indeed provides, for each credibility factor, a scale to guide the V&V activities. The available range as well as the description of the level of rigour selected have been kept faithful to those reported in the Standard. In the following, the results of the main credibility factors analysed will be presented. The main results of the computational model validation are also summarised in Table S5.

4.1. Verification: calculation verification

The discretisation error, associated with the computation of the solution at a finite number of points, was estimated with reference to the following features:

- *LH sample size*

Fig. 7 shows that as the number of falls N increases, the median, maximum and minimum (taken over the validation cohort) error computed on ARF0 decrease, becoming negligibly small ($< 0.5\%$) for $N = 10^6$ when compared to the median, maximum minimum values of ARF0 computed using $N = 10^7$.

- *Number of impact poses*

The reduced-order (surrogate) model that estimates the femoral strength as a function of the impact direction converges asymptotically to the true value (which in this case is the value predicted by the full-order FE model) as the number of samples increases. Herein, the lowest load to failure value and ARF0 estimated based on 28 and 96 full-order models produced an average error inferior to 2 and 1 percentage points (pp) respectively with respect to those obtained from 496 full-order FE models. In Fig. 8 the response surfaces obtained in the three cases are shown.

- *Mesh dimensions*

The 2 mm mesh produced an error lower than 5 pp computed on the principal strains with respect to the 0.75 mm mesh (Fig. S1 in the electronic supplementary material), which was judged an acceptable approximation error due to spatial discretisation.

- *Elasticity spatial discretisation*

The spatial discretisation of bone elasticity based on a 50 HU difference produced differences of 1.6 pp and 4 pp for the MSF and ARF0 with respect to the finest discretisation possible.

With respect to numerical solver error, the different Newton-Raphson convergence criteria did not sensitively affect the load to failure, with differences lower than 0.05 pp.

4.2. Validation: computational model

- *Governing equations form: the choice of the density to Young's modulus relationship expressions.*

The adopted density to Young's modulus relation proposed by [18] was showed to yield the best agreement between numerical calculations and experimental measurements [27].

- *System configuration form: the CT-derived femur geometry.*

The average Hausdorff distances between the femur geometries were 3.6 mm and 4.7 mm, considering the intra- and inter-subjects' variability in the segmentation. These differences affected BBCT-hip outcomes producing in both cases a coefficient of variation computed on the MSF below 5 pp and below 7 pp for the ARF0.

- *System conditions form: applied boundary conditions to simulate a fall on the side.*

In [25], the model stratification power turned out to be highest for the contact model at the greater trochanter compared to multi-point constraints and the linear models, which predicted high strains in various locations of the proximal femur including the greater trochanter, which has rarely reported previously.

- *Governing equations inputs: the coefficients of Morgan's relation to determining Young's modulus from density.*

$\rho_{app} - E$ relations based on parameters contained in the 95% confidence interval produced differences below 20 pp on the loads to failure, below 5 pp on the MSF and below 18 pp on resulting ARF0 values.

- *Governing equations inputs: the coefficients included in the HU-density calibration law to quantify voxel density from their HU.*

The different $HU - \rho_{QCT}$ relations resulting from either a wider or tighter calibration phantom segmentations based on parameters contained in the 95% confidence interval produced differences below 1 pp on the loads to failure and MSF and below 3 pp on resulting ARF0 values.

- *System configuration inputs: the locations of the anatomical landmarks identified on the femur to create the anatomical reference system used for load and boundary conditions.*

Table 1
Credibility activities and factors.

Activity	Credibility factor	Available Range	Rigour Selected	Achieved Credibility
Verification				
Code Verification	Software Quality Assurance (SQA) (5.1.1.1)	a-c	b: SQA procedures from the vendors are referenced.	Medium
	Numerical Code Verification (NCV) (5.1.1.2)	a-d	b: multiple benchmark test cases are used to verify the numerical solution.	Medium
Calculation verification	Discretisation error (5.1.2.1)	a-c	c: conservation equation balances are checked, and mesh sensitivity study conducted.	High
	Numerical solver error (5.1.2.2)	a-c	c: problem-specific sensitivity study performed on solver parameters.	High
	User error (5.1.2.3)	a-d	b: inputs and outputs verified by practitioner.	Medium
Validation				
Computational model	Model Form (5.2.1.1)	a-c	c: comprehensive evaluation of model form performed **(segmented geometry, density-elasticity relationship, principal strains-based fracture criteria, boundary conditions).**	High
	Model Inputs			
	Quantification of sensitivities (5.2.1.2.1)	a-c	c: comprehensive sensitivity analysis performed.	High
	Quantification of Uncertainties (5.2.1.2.2)	a-c	c: input uncertainties identified and propagated.	High
Comparator – Observed data	**Test samples**			
	Quantity of test samples (5.2.2.1.1)	a-c	c: statistically relevant number of samples used.	High
	Range of characteristic test samples (5.2.2.1.2)	a-d	b: samples with range of characteristics near nominal (i*n vitro* data). c: samples representing expected extreme values included (*in vivo* data).	Medium
	Measurements of test samples (5.2.2.1.3)	a-c	c: all key characteristics measured.	High
	Uncertainty of test sample measurements (5.2.2.1.4)	a-d	c: statistical treatment of repeated measurements (*in vitro* data).	Medium
	Test Condition			
	Quantity of test conditions (5.2.2.2.1)	a-c	b: two test conditions examined (*in vitro* data).	Medium
	Range of test conditions (5.2.2.2.2)	NA		
	Measurements of Test Conditions (5.2.2.2.3)	NA		
	Uncertainty of Test Conditions Measurements (5.2.2.2.4)	NA		
Assessment	Equivalency of Input parameters (5.2.3.1)	a-c	c: types and inputs equivalent (*in vitro* data).	High
	Output comparison			
	Quantity (5.2.3.2.1)	a-b	b: multiple outputs compared.	High
	Equivalency of output parameters (5.2.3.2.2)	a-c	c: types of outputs were equivalent (*in vitro* data). b: types of output were similar (*in vivo* data).	Medium
	Rigour of Output comparison (5.2.3.2.3)	a-d	b: comparison performed determining the difference between experimental and computational results. **The comparison was performed based on the Standard Error of Estimate (SEE) for *in vitro* data, Area Under Curve (AUC) for *in vivo* data.**	Medium
	Agreement of output comparison (5.2.3.2.4)	a-c	c: level of agreement satisfactory for all comparison	High
Applicability	Relevance of the Quantity of Interest (5.3.1)	a-c	a: the quantities of interest from the validation activities were related to those for the CoU (*in vitro* data) b: the quantities of interest used for the validation activities was equivalent to those for the CoU but the way it was adopted different (*in vivo* data)	Low-Medium
	Relevance of the Validation Activities on the CoU (5.3.2)	a-d	b: there was partial overlap between the ranges of the validation points and the CoU	Low-Medium

For each credibility factor, the number of the corresponding paragraph within the ASME V&V40–2018 is reported. The available range also refers to the example gradation activities reported in the guidance. The last two columns refer to the credibility level achieved. The description of the level of rigour pursued come from ASME V&V Standard. Additional parts specific to our case have been highlighted in bold. Some credibility factors are reported as Not Applicable (NA). The reason is that those factors refer to the experimental test conditions which were not relevant for the *in vitro* tests performed. The referenced experimental studies indeed tested cadaver femurs in single stance and sideways fall condition. Strains were measured with strain gauges and compared to the outcomes of the corresponding FE models.

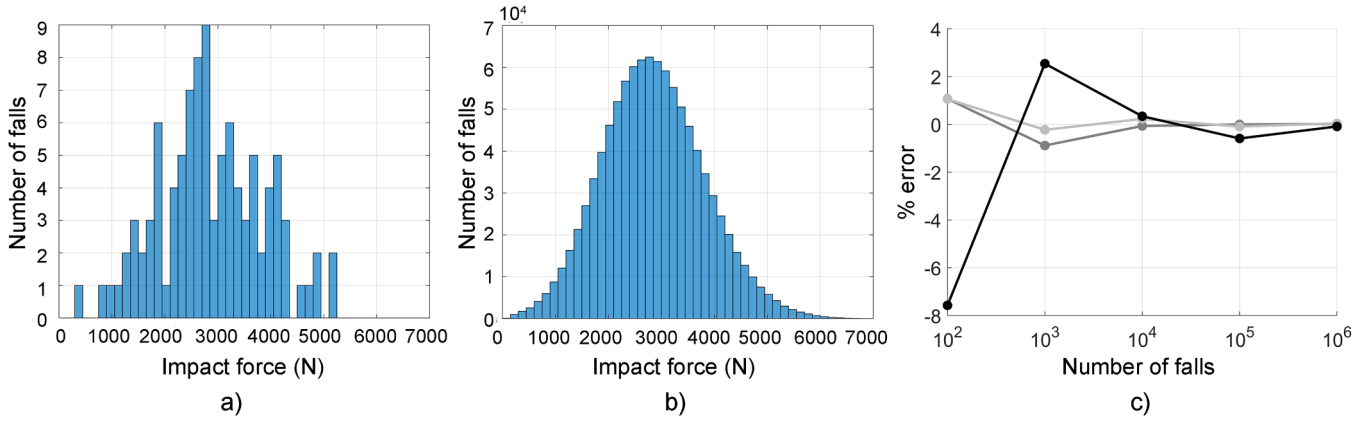

Fig. 7. The impact force values distribution as estimated by the fall mechanistic stochastic model when a) 10^2 and b) 10^6 possible falls were simulated. c) the minimum (black), median (light grey) and maximum (dark grey) error in model prediction ARF0 (expressed as percentage points, pp) over the validation cohort.

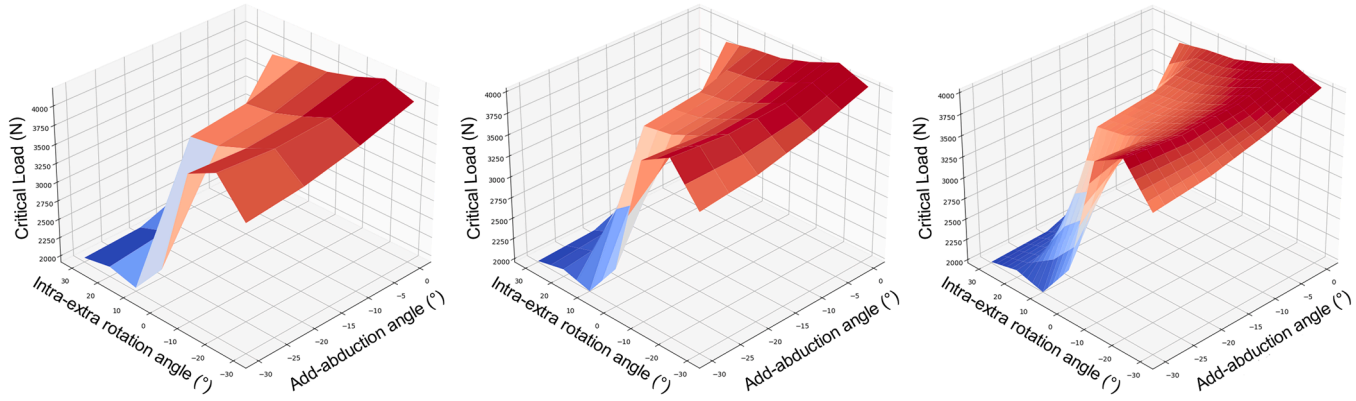

Fig. 8. The response surfaces resulting from the interpolation of progressively increasing full-order models. From left to right the surface is obtained from 28, 96 and 496 critical loads (loads to failure) computed at different impact poses in the considered intra-extra rotation and add-abduction angles range.

The coefficient of variation computed on the different identified anatomical landmarks locations (*x, y, z* coordinates) settled below 3 pp except for the mediolateral direction, where the coefficient of variation reached 20 pp. This variability slightly affected BBCT-hip outcomes, with a variation of the loads to failure below 5 pp, of the MSF below 1 pp and of ARF0 below 4 pp.

- *System conditions inputs: the portion of the distal diaphysis, computed as a percentage of the total biomechanical length, constrained to rotate around the knee centre.*

Changing the location along the diaphysis where the MPC to the knee centre was defined did not affect BBCT-hip outcomes, with variation of loads to failure, MSF and ARF0 below 1 pp.

- *System conditions inputs: the contact parameters employed in the definition of the rigid frictionless contact plane placed at the greater trochanter.*

The change in the contact parameters caused variations below 1 pp in the load to failure determined in the neutral impact pose.

The combined effect of uncertainties in the segmentation, CT calibration and anatomical landmarks identification resulted in variations below 5 pp for the load to failure and MSF, below 12 pp for ARF0.

4.3. Validation: comparator – observed data

- *Prediction accuracy*

In [22], where over 600 deformation measurements (done using strain gauges) were acquired on cadaver bones loaded both physiologically and as during a fall on a side, simulating the variability recorded *in vivo*, an error of 7 pp (root mean squared error normalised by the maximum measured deformation) was reported, confirmed also in [28].

In similar experimental settings, the cadaver femur was loaded until fracture. The failure load was predicted with an average error lower than 15 pp [29,30].

- *Stratification accuracy*

The studies on the Sheffield cohort presented in [17,16] compared ARF0 ability to separate fracture from non-fracture cases with that of aBMD-derived T-score. For ARF0 the area under the ROC curve (AUC) ranged from 0.85 to 0.87, resulting to be significantly higher than

AUC=0.75 corresponding to DXA-based T-score at the femoral neck. Only ARF0 was considered and not $\Delta ARF0$, as it was not possible to have access to cohorts with two CT scans in time available for each subject.

4.4. Applicability: relevance of the quantities of interest

The quantities of interest considered in the validation study were the strains and the load to failure for the prediction accuracy, and ARF0 for the stratification accuracy assessment. Therefore, the quantities of interest from the validation activities were considered related to those for the CoU although not identical, since the $Ave_{\Delta ARF0}$ could not be calculated.

4.5. Applicability: relevance of the validation activities to the CoU

According to the declared CoU, the average difference in ARF0 predicted by BBCT-hip between two time points is used to evaluate the (Phase II) efficacy of a new antiresorptive drug. Here, the ability of ARF0 in separating fracture from non-fracture cases was referenced. The stratification accuracy of BBCT-hip has been evaluated on a post-menopausal cohort whose distributions of body height, body mass and bone mineral density of this validation comparator reflect, by design, the distribution of osteopenia in the population. As a whole, partial overlap can be identified between the ranges of the validation points and the CoU.

5. Discussion

In silico technologies are extensively employed, in industry, during the development of new products. For instance, finite element or computational fluid dynamics models assist the design of medical devices such as plates or stents. Nevertheless, the challenge comes when modelling and simulation aim to be employed not only for research & development purposes, but also to inform a regulatory decision. In this case the question to answer does not involve a single individual anymore, as in digital twin solutions, but rather concerns the safety and efficacy of a device or intervention when used on a variety of patients. Traditionally, such question has been answered through experimentation, clinical or otherwise, but in recent years the concept of In Silico Trials appeared [1,31,32]. Before adopting an *in silico* tool for regulatory purposes, evidence is necessary to demonstrate the credibility of the predictions that the *in silico* methodology provides. Regulatory pathways change based on the class of medical product of interest, *e.g.*, a medical device or drug. As far as medical devices are concerned, FDA encourages the use of *in silico* methodologies in regulatory submissions, endorsing the technical standards ASME V&V 10-, 20- [33,34] and the most recent 40–2018. In the European Union context, the EU-MDR explicitly mentions computer modelling as one possible source of evidence in one of its annexes. However, a harmonised technical standard in the EU regulatory framework is not available yet [35]. The pharmacological field poses major challenges as it positions culturally further from *in silico* methodologies. Yet, the recently issued guideline on the reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation mirrors a rapidly changing framework.

In this context, this work aimed to present the whole workflow followed to assess the credibility of an *in silico* methodology by means of an ASME V&V40–2018-based technical validation. The process was part of a qualification advice request to EMA concerning the adoption of the BBCT methodology to test the efficacy of new treatments against osteoporosis in multidose Phase II studies. Although currently the change in the value of aBMD over time represents the primary outcome in such multi-dose Phase II studies, aBMD has demonstrated limited accuracy in hip fracture prediction. Hence, in light of the improved accuracy ARF0 has showed so far, the proposed CoU involved the use of ARF0, and in particular of ARF0 change over time averaged on the population, as the main outcome within the afore-mentioned Phase II clinical studies instead of aBMD. One could argue that both aBMD and ARF0 changes over time might be used as response variables, which could lower the regulatory impact here considered to define the model risk. However, ARF0 is determined starting from a CT-based FE model, where patient-specific densitometric information from the CT are translated into the Young's modulus assigned to each element of the mesh. A DXA image, instead, contains the subject's densitometric information projected on a plane, and averaged to extract the aBMD. This means that CT encompasses DXA, actually containing completer and more accurate densitometric information than the latter. This is the reason why in the CoU the use of the only ARF0 was proposed in replacement of the current regulatory gold standard aBMD. Therefore, the resulting burden of credibility turned out likely higher than it would be assessed otherwise, as a new replacement standard rather than a supporting standard for the regulatory decision was proffered.

Although here presented with a clear application to the pharmacological field, the technical validation procedure could be followed within broader contexts. In light of the increasing popularity of ASME V&V40–2018 adoption [5,36,37], the authors' intention was to provide a practical application of the technical standard employed to assess the credibility of an *in silico* methodology. Of course, based on the model, its CoU and risk, the goals for each credibility factor and, consequently, the rigour adopted to perform credibility activities will change, as well as the decision to consider the model credible enough once completed the credibility activities. In addition, because in this work ASME V&V 40–2018 was employed to guide the credibility assessment of an *in silico* model to be used in the drug development context, some of the original concepts were nuanced specifically, such as the use of regulatory impact instead of model influence for the risk definition. In fact, the adoption ASME V&V 40 technical standard in diversified fields witnesses its versatility, which results from its flexible framework. The reader should consider that there may be a certain degree of interpretation or judgement involved in selecting the credibility activities, as well as in defining the appropriate gradations and levels for risk and credibility. The credibility factors and the scale of possible levels of rigour each factor can be considered with are not static and need to be adapted to the specific situation. Therefore, the levels of rigour selected for the V&V activities reported in Table 1 took origin from the scale presented in the Standard, but they might then be translated to the specific context. Moreover, although ASME V&V40 mentions the possibility to use a comparator represented by both *in vitro* and *in vivo* studies, as in the here presented case, it does not provide specific recommendations on how to combine both to assess the achieved credibility level. Hence, this aspect should be tailored specifically for each case, accounting for the model, CoU and risk.

We also point out that some credibility activities were referenced as carried out in past years, rather than being performed for the specific purpose: this highlights how robust V&V and credibility activities carried out throughout the development of computational models could potentially enable future application of any such models.

Excluding material properties-related uncertainties, the variation in the quantities of interest was below 12%, which is consistent with the value identified in [38]. In addition, BBCT-hip showed good predictive and stratification accuracy. The stratification accuracy, in particular, was significantly improved with respect to the aBMD, the primary outcome currently adopted in Phase II clinical studies. Therefore, in light of BBCT-hip identified risk level and accounting for the response variable currently used in dose-response Phase II studies, we deemed BBCT-hip to be credible enough for the defined CoU. We acknowledge that the uncertainties intrinsically present in the material properties assignment affected the quantities of interest to the highest extent (~18%). Nonetheless, considering the proposed CoU, based on ARF0 change through time, that degree of uncertainty was considered acceptable. The reason is that, being ΔARF0 a relative variable (*i.e.* the difference between two ARF0 values) and being the choice of the density-Young's modulus deterministic, since that will be always kept fixed,

ΔARF0 will be only affected by the changes in the bone due to the treatment. Eventually, with respect to the proposed CoU, we also acknowledge that a fundamental part of the full credibility assessment for an *in silico* methodology, which is missing here, is clinical validation. Within the drug development field in fact, the evidence supporting the adoption of a new biomarker the regulator expects to see is represented by clinical validation, meaning a clinical study based on the construct validity, predictive accuracy, ability to detect change concepts. However, because the biomarker we propose derives from a computational model, we envisioned a so-called technical validation, which is everything we do to demonstrate our model's predictions are accurate and credible. This technical validation includes ASME V&V 40 credibility activities presented in this work, and stops at the use of retrospective clinical data (the Sheffield cohort in this case). Herein, all the validation activities, and in particular the stratification accuracy assessment, considered ARF0, due to the unavailability of cohorts where two different CT scans were taken over time. However, Δ*ARF*0 represented the main variable of interest of BBCT-hip presented context of use: in this respect, a prospective clinical validation study would be useful in establishing Δ*ARF*0 predictive capacity and sensitivity to change.

In conclusion, we have here provided the full pipeline followed to assess the credibility of BBCT *in silico* solution according to ASME V&V40–2018: it might assist and support future regulatory submissions fostering the adoption of *in silico* trials.

Open access data

The following Open Access Data are linked to this manuscript http://doi.org/10.6092/unibo/amsacta/7123

Funding

This study was supported by the European Commission through the H2020 project "In Silico World: Lowering barriers to ubiquitous adoption of In Silico Trials" (topic SC1-DTH-06-2020, grant ID 101016503).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge PRACE for awarding access to the Fenix Infrastructure resources at CINECA, which are partially funded from the European Union's Horizon 2020 research and innovation programme through the ICEI project under the grant agreement no. 800858. We also acknowledge the support of the CBM2 project, that provided computational time and expertise on scalability of the in silico trials code.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107727.

References

[1] M. Viceconti, C. Cobelli, T. Haddad, A. Himes, B. Kovatchev, M. Palmer, In silico assessment of biomedical products: the conundrum of rare but not so rare events in two case studies, Proc. Inst. Mech. Eng. H 231 (2017) 455–466, https://doi.org/10.1177/0954411917702931.

[2] L.W. Schruben, Establishing the credibility of simulations, Simulation 34 (1980) 101–105, https://doi.org/10.1177/003754978003400310.

[3] National Aeronautics and Space Administration (NASA), Standard For Models and Simulation, NASA-STD-7009, 2008.

[4] The American Society of Mechanical Engineers (ASME), Assessing Credibility of Computational Modeling Through Verification and Validation: Application to Medical Devices, ASME V&V 40-2018, 2018.

[5] T.M. Morrison, P. Hariharan, C.M. Funkhouser, P. Afshari, M. Goodin, M. Horner, Assessing computational model credibility using a risk-based framework: application to hemolysis in centrifugal blood pumps, ASAIO. J. 65 (2019) 349–360, https://doi.org/10.1097/MAT.0000000000000996.

[6] Food and Drug Administration (FDA) - Center for Devices and Radiological Health, Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions, 2021. https://www.fda.gov/media/154985/download.

[7] Food and Drug Administration (FDA) - Center for Devices and Radiological Health, Qualification of Medical Device Development Tools, 2017. https://www.fda.gov/media/87134/download.

[8] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, (n.d.).

[9] European Medicines Agency, Guideline On the Reporting of Physiologically Based Pharmacokinetic (PBPK) Modelling and Simulation, 2018. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-reporting-physiologically-based-pharmacokinetic-pbpk-modelling-simulation_en.pdf.

[10] F.T. Musuamba, I. Skottheim Rusten, R. Lesage, G. Russo, R. Bursi, L. Emili, G. Wangorsch, E. Manolis, K.E. Karlsson, A. Kulesza, E. Courcelles, J. Boissel, C. F. Rousseau, E.M. Voisin, R. Alessandrello, N. Curado, E. Dall'ara, B. Rodriguez, F. Pappalardo, L. Geris, Scientific and regulatory evaluation of mechanistic *in silico* drug and disease models in drug development: building model credibility, CPT Pharmacometr. Syst. Pharmacol. 10 (2021) 804–825, https://doi.org/10.1002/psp4.12669.

[11] C. Kuemmel, Y. Yang, X. Zhang, J. Florian, H. Zhu, M. Tegenge, S.-M. Huang, Y. Wang, T. Morrison, I. Zineh, Consideration of a credibility assessment framework in model-informed drug development: potential application to physiologically-based pharmacokinetic modeling and simulation, CPT Pharmacometr. Syst. Pharmacol. 9 (2020) 21–28, https://doi.org/10.1002/psp4.12479.

[12] European Medicines Agency, Guideline On the Evaluation of Medicinal Products in The Treatment of Primary Osteoporosis, 2006. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-evaluation-medicinal-products-treatment-primary-osteoporosis_en.pdf.

[13] A. Odén, E.V. McCloskey, H. Johansson, J.A. Kanis, Assessing the impact of osteoporosis on the burden of hip fractures, Calcif. Tissue Int. 92 (2013) 42–49, https://doi.org/10.1007/s00223-012-9666-6.

[14] S.C.E. Schuit, M. van der Klift, A.E. a. M. Weel, C.E.D.H. de Laet, H. Burger, E. Seeman, A. Hofman, A.G. Uitterlinden, J.P.T.M. van Leeuwen, H. a. P. Pols, Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam Study, Bone 34 (2004) 195–202, https://doi.org/10.1016/j.bone.2003.10.001.

[15] S.A. Wainwright, L.M. Marshall, K.E. Ensrud, J.A. Cauley, D.M. Black, T.A. Hillier, M.C. Hochberg, M.T. Vogt, E.S. Orwoll, Study of Osteoporotic Fractures Research Group, Hip fracture in women without osteoporosis, J. Clin. Endocrinol. Metab. 90 (2005) 2787–2793, https://doi.org/10.1210/jc.2004-1568.

[16] A. Aldieri, M. Terzini, A.L. Audenino, C. Bignardi, M. Paggiosi, R. Eastell, M. Viceconti, P. Bhattacharya, Personalised 3D assessment of trochanteric soft tissues improves HIP fracture classification accuracy, Ann. Biomed. Eng. 50 (2022) 303–313, https://doi.org/10.1007/s10439-022-02924-1.

[17] P. Bhattacharya, Z. Altai, M. Qasim, M. Viceconti, A multiscale model to predict current absolute risk of femoral fracture in a postmenopausal population, Biomech. Model. Mechanobiol. 18 (2019) 301–318, https://doi.org/10.1007/s10237-018-1081-0.

[18] E.F. Morgan, H.H. Bayraktar, T.M. Keaveny, Trabecular bone modulus–density relationships depend on anatomic site, J. Biomech. 36 (2003) 897–904, https://doi.org/10.1016/S0021-9290(03)00071-X.

[19] H.H. Bayraktar, E.F. Morgan, G.L. Niebur, G.E. Morris, E.K. Wong, T.M. Keaveny, Comparison of the elastic and yield properties of human femoral trabecular and cortical bone tissue, J. Biomech. 37 (2004) 27–35, https://doi.org/10.1016/s0021-9290(03)00257-4.

[20] I. Skottheim Rusten, F.T. Musuamba, Scientific and regulatory evaluation of empirical pharmacometric models: an application of the risk informed credibility assessment framework, CPT Pharmacometr. Syst. Pharmacol. 10 (2021) 1281–1296, https://doi.org/10.1002/psp4.12708.

[21] ANSYS, ANSYS Verification Manual, Release 19.3, 2019.

[22] E. Schileo, F. Taddei, A. Malandrino, L. Cristofolini, M. Viceconti, Subject-specific finite element models can accurately predict strain levels in long bones, J. Biomech. 40 (2007) 2982–2989, https://doi.org/10.1016/j.jbiomech.2007.02.010.

[23] D.R. Carter, W.C. Hayes, The compressive behavior of bone as a two-phase porous structure, JBJS 59 (1977) 954–962.

[24] T.S. Keller, Predicting the compressive mechanical behavior of bone, J. Biomech. 27 (1994) 1159–1168, https://doi.org/10.1016/0021-9290(94)90056-6.

[25] Z. Altai, M. Qasim, X. Li, M. Viceconti, The effect of boundary and loading conditions on patient classification using finite element predicted risk of fracture, Clin. Biomech. 68 (2019) 137–143, https://doi.org/10.1016/j.clinbiomech.2019.06.004.

[26] L. Cristofolini, E. Schileo, M. Juszczyk, F. Taddei, S. Martelli, M. Viceconti, Mechanical testing of bones: the positive synergy of finite–element models and *in vitro* experiments, Philos. Trans. R. Soc., A 368 (2010) 2725–2763, https://doi.org/10.1098/rsta.2010.0046.

[27] E. Schileo, F. Taddei, L. Cristofolini, M. Viceconti, Subject-specific finite element models implementing a maximum principal strain criterion are able to estimate failure risk and fracture location on human femurs tested *in vitro*, J. Biomech. 41 (2008) 356–367, https://doi.org/10.1016/j.jbiomech.2007.09.009.

[28] L. Grassi, E. Schileo, F. Taddei, L. Zani, M. Juszczyk, L. Cristofolini, M. Viceconti, Accuracy of finite element predictions in sideways load configurations for the proximal human femur, J. Biomech. 45 (2012) 394–399, https://doi.org/10.1016/j.jbiomech.2011.10.019.

[29] A. Malandrino, F. Taddei, E. Schileo, M. Juszczyk, L. Cristofolini, M. Viceconti, Prediction of failure load and location on proximal femur under a single stance loading condition, J. Biomech. Supplem. 1 (2008) S201, https://doi.org/10.1016/S0021-9290(08)70201-X.

[30] E. Schileo, L. Balistreri, L. Grassi, L. Cristofolini, F. Taddei, To what extent can linear finite element models of human femora predict failure under stance and fall loading configurations? J. Biomech. 47 (2014) 3531–3538, https://doi.org/10.1016/j.jbiomech.2014.08.024.

[31] F. Pappalardo, G. Russo, F.M. Tshinanu, M. Viceconti, In silico clinical trials: concepts and early adoptions, Brief. Bioinformat. 20 (2019) 1699–1708, https://doi.org/10.1093/bib/bby043.

[32] M. Viceconti, F. Pappalardo, B. Rodriguez, M. Horner, J. Bischoff, F. Musuamba Tshinanu, In silico trials: verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products, Methods 185 (2021) 120–127, https://doi.org/10.1016/j.ymeth.2020.01.011.

[33] The American Society of Mechanical Engineers (ASME), ASME V&V 10-2006 (R2016), Guide For Verification & Validation in Computational Solid Mechanics, 2016.

[34] The American Society of Mechanical Engineers (ASME), ASME V&V 20-2009 (R2016), Standard For Verification & Validation in Computational Fluid Dynamics and Heat Transfer, 2016.

[35] F. Pappalardo, J. Wilkinson, F. Busquet, A. Bril, M. Palmer, B. Walker, C. Curreli, G. Russo, T. Marchal, E. Toschi, R. Alessandrello, V. Costignola, I. Klingmann, M. Contin, B. Staumont, M. Woiczinski, C. Kaddick, V.D. Salvatore, A. Aldieri, L. Geris, M. Viceconti, Toward A regulatory pathway for the use of in silico trials in the CE marking of medical devices, IEEE J. Biomed. Health Inform. 26 (2022) 5282–5286, https://doi.org/10.1109/JBHI.2022.3198145.

[36] G. Luraghi, S. Bridio, C. Miller, A. Hoekstra, J.F. Rodriguez Matas, F. Migliavacca, Applicability analysis to evaluate credibility of an in silico thrombectomy procedure, J. Biomech. 126 (2021), 110631, https://doi.org/10.1016/j.jbiomech.2021.110631.

[37] M. Lopez Poncelas, L. La Barbera, J.J. Rawlinson, D. Crandall, C.E. Aubin, Credibility assessment of patient-specific biomechanical models to investigate proximal junctional failure in clinical cases with adult spine deformity using ASME V&V40 standard, Comput. Method. Biomech. Biomed. Eng. 25 (2022) 543–553, https://doi.org/10.1080/10255842.2021.1968380.

[38] F. Taddei, S. Martelli, B. Reggiani, L. Cristofolini, M. Viceconti, Finite-element modeling of bones from CT data: sensitivity to geometry and material uncertainties, IEEE Trans. Biomed. Eng. 53 (2006) 2194–2200, https://doi.org/10.1109/TBME.2006.879473.