OTA Meeting Symposium

OPEN

Artificial intelligence: international perspectives on critical issues

Meir T. Marmor, MD[a,*], Justin Krogue, MD[a], Job N. Doornberg, MD[b], Michiel Herteleer, MD, PhD[c], Adam J. Starr, MD[d], Hans-Christoph Pape, MD[e]

1. Introduction

This document presents a concise overview of the collaborative Basic Science Focus Forum (BSFF) and International Trauma Care Forum (ITCF) symposium held during the 2023 Orthopaedic Trauma Association (OTA) meeting in Seattle. The symposium's focus was "Artificial Intelligence: International Perspectives on Critical Issues." A survey conducted among the attendees indicated a unanimous consensus that the clinical adoption of artificial intelligence (AI) is inevitable within the next decade, with profound implications for clinical practice. Furthermore, a significant number of participants expressed concerns regarding the potential for AI to produce inaccurate predictions or unsound advice. They advocate for the OTA to take a central role in safeguarding orthopaedic trauma patients and surgeons against the inadvertent risks associated with AI. Discussions at the symposium spanned the current landscape of AI technology, its applications in preoperative and operative decision making, and the hurdles encountered in AI validation.

2. State of AI Technology

AI is a field of computer science that seeks to "automate intellectual tasks normally performed by humans." There are many approaches to AI. Expert systems proliferated in the 1980s and attempt to achieve AI through a hard-coded rules-based approach. However, most modern attempts at AI use machine learning (ML), in which machines are allowed to learn directly from data without hard-coded rules. In ML, data are fed into a mathematical model, which performs some computation and then outputs a prediction.

Traditional ML uses "shallow" models that involve only a few steps of computation between inputs and outputs and thus have very strong assumptions about the relationships between inputs and outputs (eg, linear and logistic regressions assume that the relationship is linear). Even if there is a true association between the inputs and output of interest, none will be found if the type of association does not match the core assumption of the model. Therefore, a critical step of traditional ML is "feature extraction," in which a human places the data into a form that is suitable for learning by the model. For example, if you were trying to predict knee function scores from x-rays with linear regression, you would first extract one or more human-coded features from the x-ray that you think may have a linear association with the function score, such as Kellgren-Lawrence grade.

Deep learning is a ML subfield that uses very "deep" models known as "neural networks" as models. Neural networks are very complex mathematical functions with many intermediate steps of computation between inputs and outputs. This complexity allows these models to be very flexible and learn most of any association that may exist. Because of this, neural networks can operate directly on raw input data without feature extraction. For example, in the abovementioned example of predicting knee function scores from x-rays, you would simply input the pixels of the x-ray as your input features.

Deep learning can currently be split into 2 basic paradigms (Fig. 1). In "Narrow AI," a deep learning model is trained on thousands of labeled examples to make predictions of interest. For example, you may train a model to recognize hip fractures from hip x-rays. With enough data, it has been consistently demonstrated that these models can achieve expert performance at a given task. These models are very useful for diagnosis and triage of course but can also be used to perform a wide variety of important tasks, including task automation (eg, automated Cobb angle measurement and assessment of femoroacetabular impingement), cognitive assistance (eg, intraoperative anatomy recognition), and many more. These models can even be used to generate new knowledge. For example, researchers at Stanford discovered additional information about knee function in OA that must be present in the x-ray by simply training a model to predict KOOS scores from x-rays and showing that it explained much more of the variability than was possible just using the typical measures of OA severity. However, there are important limitations of Narrow AI, including that it is narrow in

The authors declare that they have no conflicts of interest relating to this work, aside from: Dr. Justin Krogue—Employee and stockholder of Google LLC; Stock options in Kaliber Labs, Inc. Dr. Adam J Starr—Founding Clinical Advisor for TraumaCare AI.

[a] University of California San Francisco, San Francisco, CA, [b] University Medical Center Groningen, Groningen, Netherlands, [c] Trauma Surgery Department of the University Hospitals Leuven, Leuven, Belgium, [d] University of Texas Southwestern Medical Center, Dallas, TX, [e] University of Zurich, Zurich, Switzerland.

* Corresponding author. Address: Orthopaedic Trauma Institute, University of California San Francisco, 2540 23rd St, Building 7, 3rd Floor, Room 3907, San Francisco, CA 94110. E-mail address: Meir.Marmor@ucsf.edu (M.T. Marmor).

Source of funding: Nil.

OTAI (2025) e389

Received: 2 August 2024 / Received in final form: 19 December 2024 / Accepted: 29 January 2025
Published online 1 April 2025

http://dx.doi.org/10.1097/OI9.0000000000000389

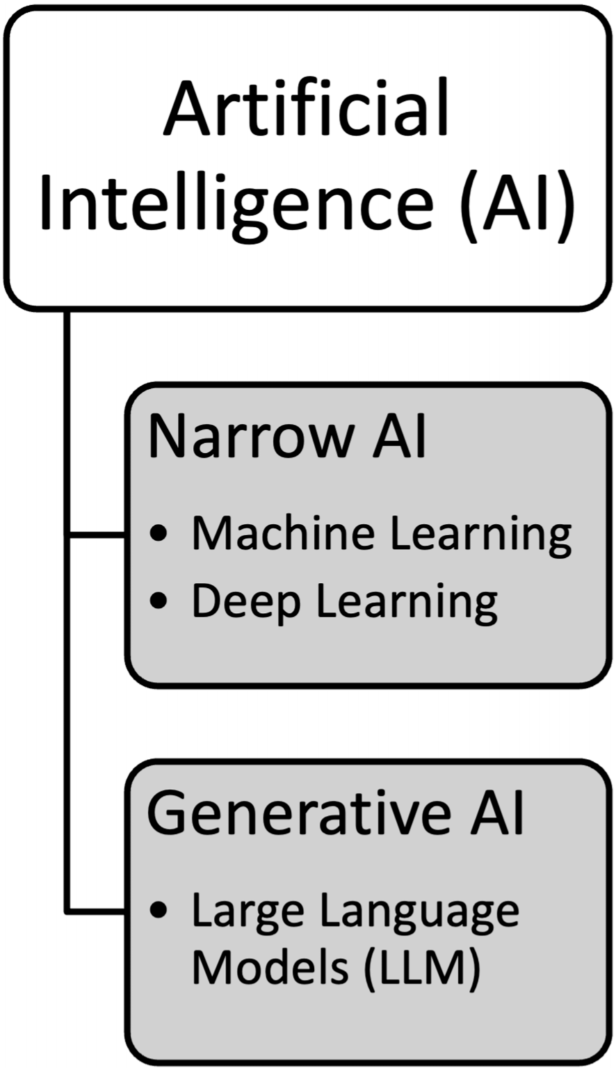

Figure 1. Artificial intelligence (AI) can be split into 2 basic paradigms: Narrow and Generative AI.

application (it has been trained to only perform 1 task and can only perform that one task); every new application requires lots of expensive, labeled data; and it often struggles to generalize to new data sets.

The other predominant paradigm is "Generative AI." This type of AI is enabled by large language models (LLMs), which are huge neural networks based on an architecture known as the transformer that are trained on huge corpuses of data to perform next sequence prediction. This is often text but can also be images (which are after all only sequences of pixels), videos, audios, or even combinations of different modalities. The large size of both these models and the data trained on has resulted in some surprising emergent features of these models, including the ability to perform well in new tasks without any significant model fine tuning. Thus, as opposed to Narrow AI, Generative AI models can be quite general purpose just by modifying the prompt submitted to the model. In addition, as mentioned above, they can be quite flexible in the type of data they work with. This has led to the development of LLMs that can perform a wide variety of tasks, such as text generation, summarization, and question answering. LLMs can also be used for medical applications, such as generating patient reports, triaging patients, and predicting outcomes. However, significant limitations of Generative AI also exist, including the problem of "hallucinations." In addition, it can be more difficult to authoritatively benchmark the performance of a LLM on a given task (ie, know how well it is expected to perform).

3. Preoperative Decision Making—Distal Radius Fractures (RAIdius Project)

To date, AI-driven computer vision research in the field of radiology has been focused on automated detection of pathoanatomy on plain radiographs, computed tomography (CT), and magnetic resonance imaging. For orthopaedic trauma, the number of deep learning models made available for fracture detection and classification is rapidly increasing, such as automated hip fracture detection as published in Lancet Digital Health.[1]

For distal radius fractures (DRFs), such AI algorithms have also been successfully trained and validated. To date, 15 studies have published deep learning algorithms to detect and/or classify DRFs on radiographs or CT, including some that are now commercially available. However, the prediction of clinical outcomes such as DRF stability has not been the focus of any studies in this field to date.

One could argue that the clinically relevant problem is not the mere presence of a DRF. The clinical challenge is to predict future loss of threshold alignment of such a DRF during follow-up, which is our main driver to offer patients surgery. An interpretable deep learning algorithm to predict the probability of loss of threshold alignment (ie, DRF instability) on plain injury radiographs holds the promise to be a game changer in clinical care, as it will eliminate human (surgeon) biases and thus augment patient-centered shared surgical decision making.

In general, treatment of DRFs is either conservative with a cast or surgical, most often by open reduction and internal fixation (ORIF) with plates and screws. Treatment decisions are distinct per country, hospital, or even surgeon: "what you get depends on where you live, and who you see." In Australia, patients are offered surgery in up to 80% of DRFs, while in the United States and Netherlands, operative rates are 26% and 10%, respectively. One could argue that such variation is undesired: some patients may be overtreated while some may be undertreated.

Shared surgical decision making in an early treatment phase remains challenging because it is based on surgeons' "art" to deem a fracture stable or unstable. However, studies have shown that human estimation of the probability of future loss of threshold alignment of a DRF is fallible. An online experiment by the Science of Variation Group (Boston, MA) showed that surgeons accurately predicted redisplacement of a DRF based on plain injury radiographs in only 54% of patients with a reduced DRF, a guesstimation. In a subsequent experiment, surgeons evaluated both radiographs and injury CT scans of reduced DRFs: accuracy improved up to 70% with the use of CT.

In short, the diagnostic performance characteristics of human interpretation of injury radiographs to determine fracture stability of DRFs may be unsatisfactory for clinical practice. For this reason, a reliable and interpretable deep learning model to predict loss of threshold alignment in DRFs, with a higher accuracy than surgeons, is of significant interest to reduce undesired treatment variation. Moreover, this approach may also shift the interest in the field of AI-driven computer vision research from the rather simple detection of pathoanatomy to prediction of clinical outcomes.

Therefore, we included patients with DRFs from retrospective databases from 2 hospitals and one multicenter prospective database. Patients were required to be primarily treated with a cast, have complete radiological follow-up, and not have received surgery while still "stable." Trauma and reduction radiographs, both posteroanterior (PA) and lateral, were collected when available, as well as sex and age. To augment training of the algorithm, 21 landmarks (11 on PA radiographs, 10 on lateral) were manually annotated on radiographs. A convolutional neural network (CNN) was trained on these radiographs and annotations.

The algorithm was trained on 2136 radiographs (583 cases) and tested on 563 radiographs (150 cases), taking 50 patients from each cohort. Our model had an AUC of 0.83 and achieved 76% accuracy, 84% sensitivity, and 68% specificity in predicting future DRF loss of threshold alignment.

In contrast to previous deep learning models to detect pathoanatomy, these results are promising because this is the first CNN algorithm used in (orthopaedic trauma) surgery to make a prediction about future fracture alignment.

With the rise of artificial intelligence (AI), in particular, deep learning methodologies with CNNs that can analyze images such as plain radiographs, these algorithms will aid surgeons in objectively quantifying the probability of loss of threshold alignment of DRFs.

4. Intraoperative Decision Making Using Augmented Reality—Use in Placement of Iliosacral Screws

Iliosacral screw placement is considered a complex procedure that requires a thorough understanding of the patient's sacral anatomy. The presence of sacral dysmorphism, osteoporotic bone, obesity, and overlying bowel gases makes it even harder to visualize the osseous fixation pathways on conventional perioperative fluoroscopy.

Owing to the need for highly skilled surgeons, the large increase in osteoporotic pelvic fractures, and the benefit that many of these patients have from minimally invasive fixation, it is becoming more difficult to provide the appropriate care in time.

Navigation can be used to simplify this procedure so that the complex sacral anatomy is no longer a threshold for the less experienced surgeon. These navigation systems, however, are very expensive and require some form of intraoperative CT imaging. This cost must be reduced and the necessary hardware simplified to make this a safe procedure for less experienced hands.

Mixed reality (MR), augmented reality (AR), and extended reality (XR) overlap and are used to describe an environment where real-world and computer-generated contents overlap so that virtual and real objects can interact in real time. This overlap is usually created using augmented reality glasses (headset) such as the Microsoft Hololens, Google Glass, or the Apple Vision Pro. This means, for example, that a patient's deep anatomy that has been previously modeled using, for example, a CT scan can be overlaid onto the patient's skin. This results in an improved visualization and better understanding of the anatomy, and it can also be used to plan and assess a surgical procedure.

The AR workflow that was developed for the planning and placement of iliosacral screws is shown in Fig. 2. Preoperatively, a CT scan is performed to assess the fracture (step 1). During the CT scan, skin markers are attached to the patient. These are later used to register the 3D models for the patient. With the help of artificial intelligence, the segmentation (creation of 3D models from CT images) of the patient's anatomy and the planning of the screw trajectories occur in an automated way (step 2). This drastically speeds up this process. The planned screw trajectories can then be assessed and adapted in a multiuser viewer application (step 3). This can be performed in both an augmented or a virtual environment. In the last phase, the 3D models and planned screw trajectories are projected onto the patient in the operating theater. The intraoperative fluoroscopy is used to verify the chosen trajectories.

Surgical augmented reality applications and surgical navigation are complementary tools. They can be used combined or separately. The main challenge for both technologies is an adequate registration of the planned trajectories or projected models. If the registration is poor, the accuracy will be poor. The improvement of this registration accuracy and turning the registration into a robust, fail-safe procedure are key for this technology to succeed and potentially reduce the costs associated with it. Artificial intelligence can help us to improve this and make the registration more software-based than hardware-based. Currently, intraoperative CT remains of importance. The hardware costs will also drop when an accurate registration is possible without intraoperative CT.

In conclusion, augmented reality can help us visualize complex anatomy, but technical improvements still need to be made to improve accuracy. A robust registration method without intraoperative CT is potentially a game changer for cost reduction. Artificial intelligence can assist with this registration and automate the workflow, making it fast and easy to use.

5. AI for Outcome Prediction: Watson Health Trauma Pathway Explorer—A Teaching Tool for Polytraumatized Patients by Using Visual Analytics

Polytraumatized patients face a spectrum of complications and adverse outcomes that preexisting conditions, comorbidities, specific injury patterns, and pathogenetic changes can influence. This complex landscape underscores the critical need for early assessment to identify risk factors of complications, which is essential for patient care and educating medical residents and young attending physicians.

In response to this need, a new visual analytics tool was developed in collaboration with IBM over several years. This tool was designed to predict special risk situations for complications using data from patient records following approval from the local institutional review board and a trauma database at a Level I trauma center. The inclusion criteria for this study were patients older than 16 years with an Injury Severity Score (ISS) greater than 16.[2]

Throughout the development phase, parameters associated with the development of complications were meticulously assessed. These parameters included patient age, abbreviated injury scales, ISS, ATLS shock classification, surgical strategy decisions, and various physiological measurements on admission. The daily monitoring of various laboratory values and the transfusion volumes was also documented. The primary end points were early in-hospital mortality (within 72 hours), sepsis, and SIRS.

IBM's Watson was used for the pathway development. Watson is a cutting-edge AI program that created projections of the clinical course, which were then visualized using a Sankey analytics tool. Sankey diagrams, named after Irish Captain Matthew Henry Phineas Riall Sankey, visually represent flow data, where the width of each flow pictured is proportional to the volume (Fig. 3a and b).

The initial data set for the first stage of our project included 3655 patients. The final data set included 1925 patients after

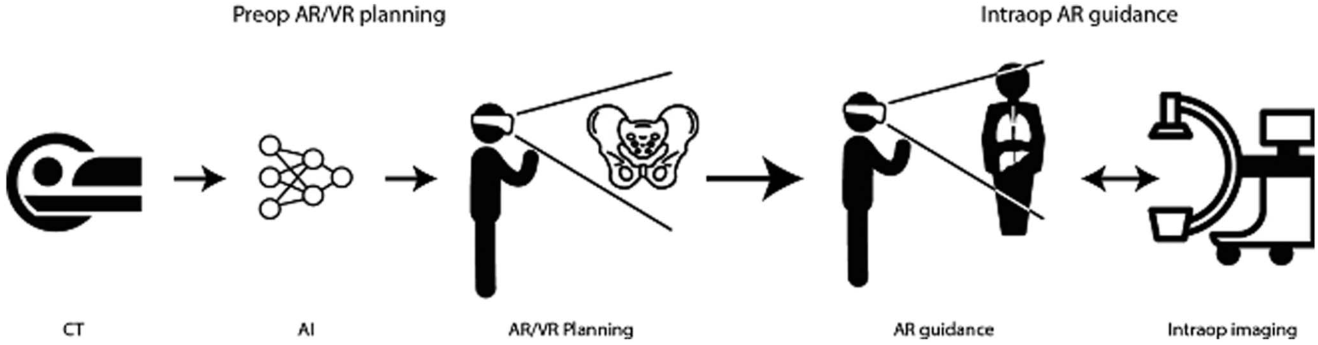

Figure 2. Augmented reality (AR) workflow for planning and placement of iliosacral screws.

stratification and the exclusion of 1730 due to incomplete data. We leveraged the Watson Health Trauma Pathway Explorer to display parameters, individual patient data, surgical pathways, ATLS groups of hemorrhage, and outcomes. Each aspect could be modified according to individual data sets and the desired outcomes. These interactive Sankey diagrams allowed users to navigate clinical pathways based on real-world data and make informed decisions in critical situations. This tool provides a sophisticated, personalized, evidence-based trauma care option by incorporating a learning component into knowledge-based artificial intelligence.

An illustrative scenario showcased the impact of age on the risk of sepsis. Patients older than 65 years had a 33% risk of developing sepsis while patients aged between 29 and 65 years had a risk as low as 2% (Fig. 3a and b). The power of visual analytics lies in its ability to unite the computational strength of machines with the perceptual and cognitive capabilities of humans. It serves as a medium between users, the analytic system, and data, enabling the exploration of raw data, extraction of insights, hypothesis generation, scenario simulation, and, ultimately, the acquisition of new domain knowledge. The true value of visual analytics is assessed in the context of the domain and the data's applicability to the user's tasks. It can shorten response time, reveal new insights, distill the essence of the data, and build trust in the analysis.

Visual Pathway Analytics enables the construction of data-driven models for clinical pathways at the cohort level. It tackles the challenges posed by the high dimensionality and variance of data, often captured in operational systems that are mere proxies for the required analysis data. This tool, equipped with a data ingestion and transformation engine, uses algorithms for medical code grouping, statistical frequency grouping, time interval simplification, and other necessary simplifications. Users can interactively explore the summarized pathways in a node-link flow diagram, a Sankey diagram variation, which allows for the investigation of treatment distributions and the identification of pathways leading to favorable or unfavorable outcomes.

Overall, the Watson Health Trauma Pathway Explorer represents an innovative step toward integrating knowledge-based AI.

6. AI for Outcome Prediction: Parkland Trauma Index of Mortality—Real-Time Predictive Model for Patients With Trauma

Vital signs and certain laboratory values have long been recognized by trauma surgeons as reliable indicators of mortality risk. Researchers have devised various scoring systems to identify patients at the highest risk of death. Such systems are invaluable for guiding clinical decisions, determining patient stability for definitive surgery, the necessity for damage control measures, and further resuscitation requirements. However, the practical application of these tools presents challenges; they typically necessitate manual calculations and lack clarity regarding the timing of these calculations. Consequently, these tools have not been widely adopted despite relying on variables known to be critical.

The Parkland Trauma Index of Mortality (PTIM) was developed to address these issues. It is a machine learning algorithm that uses data already collected in patients' electronic medical records, negating the need for surgeons to manually input data. This algorithm functions within the Epic system, providing mortality predictions for the initial 48 hours of a three-day hospital stay, with updates every hour to reflect changes in patient physiology. The PTIM model was developed with data from 1935 trauma patient encounters between 2009 and 2014 and was subsequently validated with 516 patient encounters from 2015 to 2016. Its performance was retrospectively evaluated after 1 year of clinical use.[3]

The accuracy of the model has proven to be high. In its first year of clinical use involving 776 patients, the PTIM's final score accurately predicted 20 of 23 twelve-hour time intervals leading to mortality, with a sensitivity of 86.9% (95% confidence interval [CI], 73%–100%) and a specificity of 94.7% (95% CI, 93%–96%). The positive predictive value was reported at 33.3% (95% CI, 21.4%–45%) while the model successfully predicted survival in 716 time intervals with only 3 errors, yielding a negative predictive value of 99.6% (95% CI, 99.1%–100%). The PTIM's performance, as measured by the area under the receiver operating characteristic curve, was an impressive 0.97, significantly surpassing models based on singular variables such as lactate or base deficit.

By using standard variables in an automated, hourly updated model, the PTIM addresses many of the shortcomings found in earlier models. It can potentially enhance decision making for trauma patients early in their hospitalization. Current efforts are focused on collaborating with other medical centers to validate the PTIM and broaden its implementation across additional institutions.

7. Discussion

Integrating artificial intelligence (AI) into clinical practice is expected to occur within the next decade, significantly affecting patient care. The symposium showcased that AI algorithms could

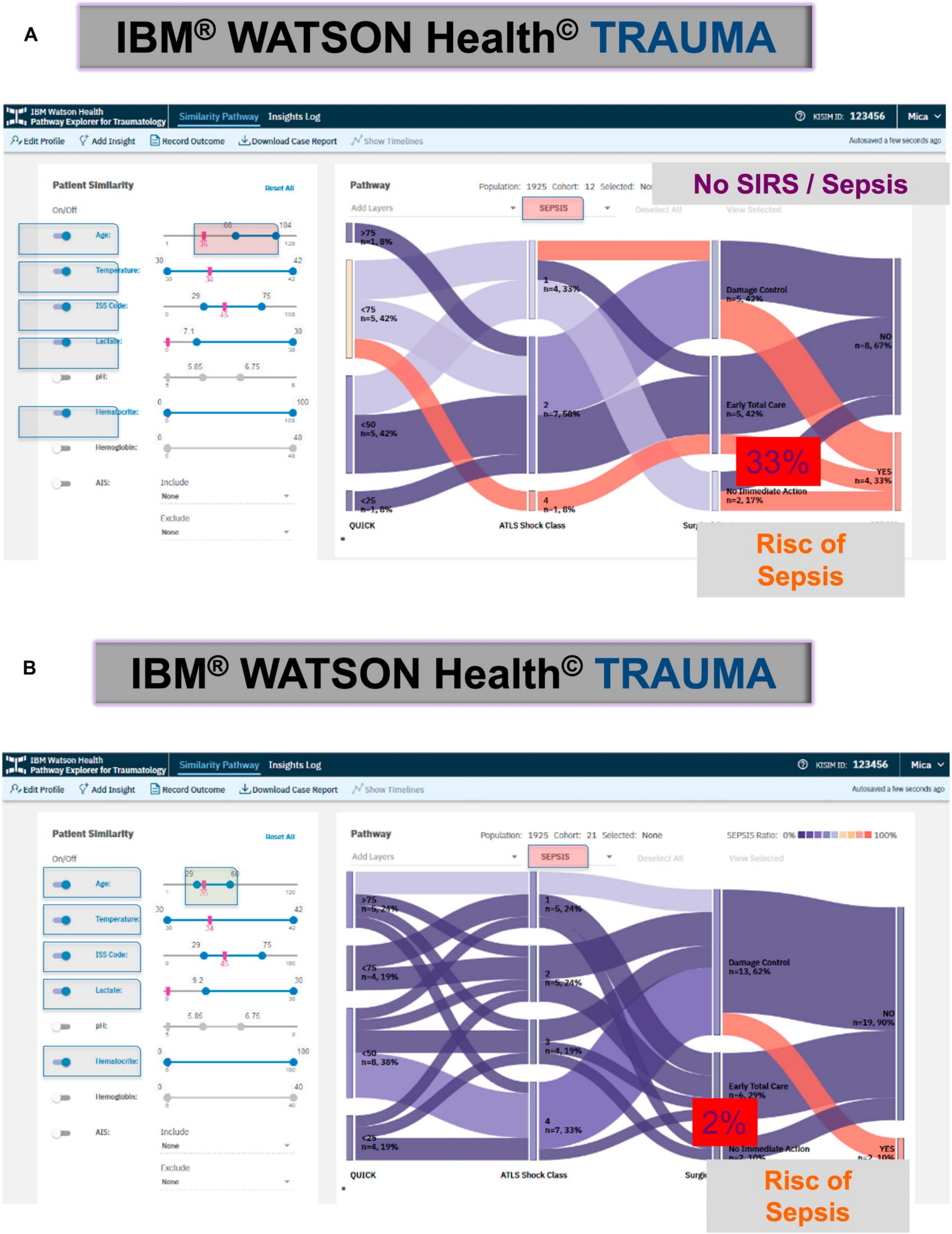

Figure 3. An illustrative scenario showcasing the impact of age on the risk of sepsis. Patients older than 65 had a 33% risk (A) of developing sepsis while patients aged between 29 and 65 years had a risk as low as 2% (B).

be incorporated at each stage of clinical treatment. Discussed applications, representative of Narrow AI, span preoperative diagnostics and fracture evaluation to surgical indication determinations, intraoperative support for surgical execution, and postoperative forecasts of complications and mortality. Advancements in Generative AI are expected to enhance AI's capabilities further, introducing sophisticated tasks such as virtual patient assistance, automated documentation, and automated preoperative planning.

It falls to professional organizations such as the Orthopaedic Trauma Association (OTA) to facilitate the seamless integration of AI into clinical settings by offering support to developers of AI technology in industry and academia. OTA interventions might include the establishment of a specialized orthopaedic trauma data repository for AI developers, creating practice guidelines, and setting performance benchmarks for new models that can be subjected to rigorous audits.

References

1. Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health*. 2022;4:e351–e358.
2. Mica L, Niggli C, Bak P, et al. Development of a visual analytics tool for polytrauma patients: proof of concept for a new assessment tool using a multiple layer Sankey diagram in a single-center database. *World J Surg*. 2020;44:764–772.
3. Starr AJ, Julka M, Nethi A, et al. Parkland trauma Index of mortality: real-time predictive model for trauma patients. *J Orthop Trauma*. 2022;36: 280–286.