Progress of Digital Technologies in the Perioperative Management of Postoperative Delirium Among Older Adults

Ann. Ital. Chir., 2026 97, 9: 1561–1574
https://doi.org/10.62713/aic.4800

Yao Lin[1], Tianmin Lu[2], Chunlan Liu[3], Xiuyun Wu[2], Haiyan Jiang[2], Mingli Zhu[2], Weiying Dai[4]

[1]Department of Nursing, Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, 310006 Hangzhou, Zhejiang, China
[2]Department of Nursing, Affiliated Hangzhou First People's Hospital Chengbei Campus, School of Medicine, Westlake University, 310022 Hangzhou, Zhejiang, China
[3]Department of Geriatrics, Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, 310006 Hangzhou, Zhejiang, China
[4]Department of Intensive Care Unit, Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, 310006 Hangzhou, Zhejiang, China

Postoperative delirium (POD) is one of the most common and severe complications among older adults, and is closely associated with poor outcomes, increased healthcare costs, and long-term cognitive decline. With the rapid growth of the global aging population and the continuous emergence of new geriatric surgical procedures, the prevention and management of POD have become a major challenge. Traditional risk assessment tools have inherent limitations in clinical practice and are insufficient to meet the growing demands of precision medicine. In recent years, digital technologies have shown tremendous potential in perioperative management; innovations such as machine learning–based predictive models, wearable sensors, electroencephalogram (EEG) monitoring, and large language models are increasingly being adopted for POD risk identification. This review comprehensively summarizes research advances in digital technology for POD risk identification and their role in clinical decision support for older adults, with the aim of providing evidence-based insights and implementation guidance for clinical practice.

Keywords: postoperative delirium; older adults; digital technology; artificial intelligence

Introduction

Owing to the accelerated aging of the global population, the number of older adults undergoing surgical procedures has increased substantially, with approximately one-third of all surgeries performed in individuals aged 65 and older [1]. In China, the proportion of patients aged 60 and older undergoing total knee replacement has increased from 79.1% to 85.6% between 2013 and 2019 [2]. The increasing surgeries among the older population poses unique challenges for perioperative care in geriatrics, as the vulnerability resulting from reduced physiological reserves in older adults elevates the risk of various postoperative complications [1]. Therefore, failure to effectively prevent and control postoperative complications will adversely affect the prognosis of older adults [3].

As one of the most common and severe postoperative complications among older adults, postoperative delirium (POD) not only prolongs hospital stays, reduces quality of life, and increases hospitalization costs, but may also lead to long-term cognitive impairment, functional dependence, and even death. POD typically occurs within the first 1–3 days after surgery and is characterized by confusion, impaired cognition, and memory loss [4]. The incidence of POD in the general surgical population is only 2.5–3%, but this rate surges to 10–20% in individuals aged 60 and older [5]. In the intensive care unit (ICU), the incidence of delirium in older adults can exceed 80%, particularly following prolonged mechanical ventilation or sedation [4]. Furthermore, delayed diagnosis is relatively common in geriatric POD. It has been statistically reported that up to 67.6% of POD cases in older adults present as hypoactive delirium (characterized by silence, apathy, sluggishness, and decreased attention), which is easily overlooked in clinical practice [4]. Furthermore, when patients already have cognitive impairment preoperatively, the acute and fluctuating symptoms of delirium are often misinterpreted as underlying cognitive impairment or residual effects of anesthesia, rather than a new episode of delirium. Despite the absence of effective treatments, POD is generally preventable, underscoring the need to develop predictive models to allow rapid identification of surgical older adults at high risk for POD. However, existing delirium risk assessment tools fail to meet the demands of personalized precision medicine. In

Submitted: 27 May 2026 Revised: 29 June 2026 Accepted: 7 July 2026 Published: 15 September 2026
Correspondence to: Mingli Zhu, Department of Nursing, Affiliated Hangzhou First People's Hospital Chengbei Campus, School of Medicine, Westlake University, 310022 Hangzhou, Zhejiang, China (e-mail: hzgfgf2007@163.com); Weiying Dai, Department of Intensive Care Unit, Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, 310006 Hangzhou, Zhejiang, China (e-mail: dwy13867448801@163.com).
Editor: Ismail Zaed

addition, these tools largely rely on clinicians' subjective judgment, take a long time during the assessment process, and have insufficient sensitivity for identifying early-stage or mild delirium.

With the rapid advancement of digital health technologies, digital tools are playing an increasingly important role in perioperative risk identification and clinical decision-making. Currently, digital health technologies are widely used for complication prediction and personalized risk stratification (preoperative), real-time adjustment of anesthetic depth (intraoperative), and continuous monitoring of behavior and physiological functions via wearable devices (perioperative), thereby establishing a comprehensive digital management framework [6,7,8]. A recently published systematic review firmly supports that digital technologies can enhance the prediction, diagnosis, and management of delirium across all populations [6].

However, due to age-related declines in physiological reserves, functional decline and cognitive impairment resulting from POD in older patients are often irreversible and can markedly increase the long-term risk of dementia [9]. Consequently, compared to other age groups, the perioperative management in older adults presents unique challenges. Nevertheless, comprehensive reviews focusing on digital management of POD in older adults, who are generally characterized as the high-risk population for POD, remain relatively scarce.

From the perspective of model development, the digital methods discussed in this review—including machine learning–based predictive models, electronic health record (EHR)–based tools, metabolomics, wearable sensors, electroencephalogram (EEG) monitoring, and large language models—share a common core: prediction. Specifically, these methods assess an individual's risk of developing POD based on multimodal data. The predictive nature of these methods enables different forms of clinical applications: (i) risk prediction (preoperative estimation of the probability of POD at the individual level); (ii) screening (rapid identification of potentially existing POD in a broader population); (iii) monitoring (continuous tracking of consciousness, activity patterns, or physiological parameters during the perioperative period); and (iv) clinical decision support (providing actionable recommendations or intervention triggers based on predictive results). Although these applications share common underlying algorithmic principles, the selection of digital tools, data sources, and implementation strategies should be tailored to specific clinical tasks and perioperative time points. In this review, we systematically examine each of these digital approaches with a focus on their respective roles, evidence maturity, and the barriers to clinical translation in the context of POD risk identification in older surgical adults, aiming to provide a practical reference framework for clinicians.

Current Applications and Limitations of Traditional Risk Assessment

Traditional POD risk assessment primarily relies on clinical scoring systems and scales. Preoperatively, the focus is on assessing risks associated with delirium, including nutritional status, sleep quality, pain, medication use, functional status, frailty, cognitive function, anxiety, depression, and history of delirium [10]. Postoperatively, non-psychiatric healthcare professionals mainly use the Confusion Assessment Method (CAM) and its derivative versions (the intensive care unit Confusion Assessment Method [CAM-ICU] and the Intensive Care Delirium Screening Checklist [ICDSC]), as well as the Nursing Delirium Screening scale (Nu-DESC), to assess delirium. Although these tools are widely used in routine practice, they fail to meet the demands of individualized precision medicine.

A systematic review and meta-analysis demonstrated that the sensitivities of Nu-DESC and CAM for postoperative delirium were only 0.69 and 0.65, respectively, indicating suboptimal diagnostic performance in identifying early-stage or mild delirium [11]. Furthermore, the assessment of these tools is highly operator-dependent, leading to a high probability of inter-rater variability [12]. Furthermore, in real-world settings where high-intensity workflows are common, clinicians' time constraints and cognitive load often result in missed diagnoses or delayed diagnoses, which deepen the challenges in capturing cognitive fluctuations of patients during the subclinical stage. Similarly, intermittent bedside delirium assessments (based on traditional measurement tools) cannot provide continuous and dynamic monitoring, making it challenging to perform individualized risk stratification.

Traditional assessment tools are unable to achieve individualized risk stratification, largely due to the following perioperative factors: surgical trauma, depth of anesthesia, intraoperative hemodynamic fluctuations, postoperative pain, and metabolic disturbances, all of which have a substantial impact on the development of POD [13]. Traditional risk scoring systems are typically based on limited clinical variables and struggle to integrate multidimensional, multimodal perioperative data (including laboratory test results, imaging features, and data from patients' previous health records). Moreover, variables in traditional measurement tools are often processed by linear weighting methods, which fail to identify complex interactions and nonlinear relationships among variables. Thus, the accuracy of risk stratification by traditional measurement tools is generally suboptimal.

Key limitations of these tools are not merely operational (that is, they cannot be fully resolved through more rigorous training or stricter adherence to operating procedures). Rather, they are structural: traditional tools are intermittent and snapshot-based, with fixed panels and reliance on subjective judgment, which render them fundamentally incompatible with the core requirements of precision medicine—namely, the emphasis on continuous, personalized, and quantifiable risk assessment. This structural gap highlights the need to adopt complementary digital approaches.

Despite these limitations, it must be acknowledged that CAM, CAM-ICU, ICDSC, and Nu-DESC remain crucial clinical reference tools, providing standardized, bedside-usable frameworks for delirium identification. Their widespread use, low cost, and lack of reliance on specialized equipment render them indispensable in routine clinical practice, particularly in resource-limited settings. Therefore, digital technologies should be viewed as a complement to specific structural gaps in traditional assessments—particularly in terms of continuous monitoring, objective quantification, and automation—rather than as a substitute for bedside clinical judgment. In fact, the optimal perioperative strategy likely lies in the synergistic integration of both, where traditional tools provide clinical rationale and interpretive context, while digital tools expand the temporal dimension and scope of quantitative analysis. In the following sections, we will explore how various digital approaches complement (rather than replace) the established role of standardized bedside assessments in identifying the risk of POD.

Predictive Modeling for Delirium Risk

Performance Trade-offs Between Machine Learning Algorithms and Traditional Regression

In the field of clinical prediction, the term 'traditional regression' generally refers to the statistical regression approaches (that is, logistic regression, linear regression, and Cox proportional hazards regression), which prioritize interpretability and hypothesis testing (focusing on estimating odds ratios, hazard ratios, and confidence intervals to understand the relationships between variables). In contrast, machine learning (ML) algorithms (such as random forests, gradient boosting) prioritize predictive accuracy and generalization performance over interpretability. These two paradigms differ fundamentally in their objectives, assumptions, and methodological frameworks.

In recent years, ML-based algorithms have demonstrated unique advantages in delirium prediction, as they can overcome the inherent constraints of traditional regression. Traditional regression requires defining variables and their functional forms within the model and assumes a linear relationship between the dependent and independent variables (though nonlinear relationships can be addressed through specific methods). However, it struggles to capture complex nonlinear relationships and is prone to overfitting when handling high-dimensional data [14]. In contrast, ML and deep learning–based algorithms (including random forests, extreme gradient boosting, and neural networks) do not require predefined assumptions about the relationships between variables. By integrating multidimensional variables, these algorithms outperform traditional regression methods when handling nonlinear relationships in high-dimensional data. Therefore, ML-based models have shown promising performance in predicting POD. Automated data extraction from EHR holds promise for further enabling real-time preoperative risk identification. Under the context of integrating automated technologies with hospital health information systems, artificial intelligence (AI) approaches, such as natural language processing and structured data mining, can be leveraged to extract risk features from previous EHR, assessment records at baseline, and real-time physiological data. This automated extraction would enable real-time updates of risk predictions, providing a more rational basis for the targeted allocation of prevention resources.

However, the advantages of ML over traditional regression are not evident in all scenarios. Although ML theoretically offers greater data utilization capabilities, such superiority does not hold true for the specific task of predicting POD. Penalized logistic regression models perform on par with ML methods in predicting POD [15]. Similarly, an EHR-based POD prediction study found that XGBoost did not perform significantly better than traditional logistic regression models (LASSO regression) in predicting POD among patients with an average age of 68.5 years [15]. In a study conducted on older adults to predict the development of POD, the performance of traditional logistic regression was compared with that of five ML models (random forest, GBM, AdaBoost, XGBoost, and stacking ensemble model). It was found that traditional logistic regression had the best area under the receiver operating characteristic curve (AUC) (0.783) and sensitivity (74.2%), although with the lowest predictive accuracy [16].

It is reasonable that complex ML algorithms do not consistently outperform traditional logistic regression in POD prediction. First, sample size plays a critical role. Generally, the theoretical advantages of ML algorithms, especially in the capacity to capture high-dimensional interactions and nonlinear patterns, cannot be fully realized in a sample size of typically hundreds to a few thousand cases. With such a sample size, the greater flexibility of ML models increases the risk of overfitting, whereas logistic regression often confers better generalizability due to parsimony [17,18]. Second, data structure matters. When the predictors are predominantly numerical clinical variables in routine practice (such as age, laboratory test results, comorbidity indices), their underlying relationship with POD is often linear or involves low-order interaction; logistic regression is sufficient to model such relationships, and the marginal gains offered by ML in these settings are inherently limited [17,19]. Moreover, the AUC alone may obscure the true comparative value of different modeling approaches [20,21,22].

Based on the aforementioned considerations, we recommend a hierarchical strategy for developing POD prediction models. In most clinical scenarios where predictors are limited to conventional structured variables, traditional regression should be used as the baseline and primary modeling method. Machine learning methods should be considered only when specific conditions are met: (i) the sample size is sufficient to support the complexity of the selected algorithm; and (ii) the data include high-dimensional or unstructured data sources (such as EEG time-series features, metabolomic spectra, or continuous wearable sensor data), in which nonlinear interactions are theoretically more pronounced. Most importantly, when comparing multiple modeling strategies, we advocate for an evaluation framework that extends beyond AUC-grounded assessments—priority should be given to calibration and decision clinical analysis as the ultimate criteria for assessing a model's practicality, as these metrics directly reflect whether improvements in predictive accuracy translate into better clinical decision-making. General recommendations for developing and evaluating (further applying) the predictive model have been provided in detail elsewhere [23].

Feasibility of Automated Risk Stratification Based on EHR

Machine learning–based predictive models constructed with EHR data have proven the technical feasibility of risk stratification based solely on routine clinical data. The SURGE-Ahead project developed a linear support vector machine–based model to predict POD, incorporating only 15 clinical and demographic features—such as age, American Society of Anesthesiologists (ASA) classification, estimated glomerular filtration rate, and recent history of falls—that can be automatically extracted from routine EHRs. In an external validation cohort (comprising 173 older adults), the model achieved an AUC of 0.86 with good calibration [24]. Another study retrospectively extracted data from middle-aged and older adults in the EHRs of the Indiana University Health System, with only 143 features routinely collected at admission. Models were constructed using multiple methods (including traditional regression and ML-based models), among which the POD prediction model built with extreme gradient boosting performed best (AUC = 0.79) [25]. Feature importance analysis in this study consistently identified ASA classification as the most important predictor across the entire set of variables in all models, validation datasets, and surveillance periods, while the ranking of other features (such as the surgical specialty, emergency surgery status, and medication use) varied across institutions and surveillance periods [25]. Similarly, a recent study focusing specifically on older adults undergoing femoral neck fracture surgery under spinal anesthesia developed a nomogram based on eight independent risk factors derived from routine clinical data (including age, diabetes, preoperative hemoglobin and albumin, ASA grade, and postoperative pain scores), demonstrating a good discriminative ability for POD prediction [26]. However, these studies were conducted in an offline setting where data extraction and modeling were performed by researchers, and automated deployment by integrating into the health information systems has not yet been achieved.

However, it is important to recognize that the aforementioned study was conducted in an offline, retrospective setting where data extraction, preprocessing, and modeling were all performed by researchers retrospectively rather than in real time [27]. While showing technical feasibility, the retrospective and offline design represents only the initial stage in the evidence hierarchy for clinical predictive models [27]. Several additional validation steps are required between retrospective model development and real-time clinical application: (i) prospective temporal validation to ensure that model performance remains stable over the long term as clinical practice and patient populations evolve; (ii) external validation across different institutions and geographic settings to assess the model's generalizability; and (iii) most importantly, prospective study of clinical impact to evaluate whether the risk alerts generated by the model actually translate into better clinical decisions and improved patient outcomes [28]. To date, no EHR–based POD prediction model has been prospectively deployed into real-time clinical workflows and evaluated for its impact on patient care [15]. When interpreting the existing evidence, this gap between retrospective technical validation and prospective clinical utility should be explicitly acknowledged [27].

If a fully validated predictive model is integrated into a hospital's health information system via a specific algorithm, the system can automatically perform feature extraction, data preprocessing, and risk scoring upon a patient's admission, without requiring additional intervention from clinicians. The implementation of this automated model would transform delirium risk stratification from a static, single-point-in-time assessment into an evaluation with a dynamic trajectory, better aligning with the multifactorial and time-sensitive nature of POD risk. Through automated workflows, it is expected to expand the coverage of risk assessment (in terms of both the number of patients assessed and the timing of assessments), the degree of data standardization, and the timeliness of results. An automated stratification framework based on EHRs is a critical approach to transitioning toward objective, continuous, and data-driven systematic assessment.

Precise Prediction Through the Integration of Preoperative Multimodal Metabolomic Data

Key Decision Support Requires High-Precision Prediction

The conventional belief holds that predictive models should be built with inexpensive, widely available, and routinely collected indicators as predictors to ensure the models' deployability and generalizability in the real world (that is, achieving risk stratification at minimal additional cost). However, due to the critical time-sensitivity of POD identification—where delayed detection may lead to irreversible adverse outcomes—the demand for model accuracy in POD prediction is substantially heightened. In addition, preoperative POD risk stratification is commonly used to inform complex clinical decisions and the allocation of medical resources, including whether to initiate intensive, multidisciplinary preoperative prehabilitation and delay surgery, as well as the selection of intraoperative anesthesia regimens. Since imprecise predictions may misguide clinical decisions and increase the risk of other complications, high-accuracy preoperative and intraoperative prediction of POD is essential. However, the aforementioned low-cost predictive models may struggle to capture neurobiological vulnerability at the individual level, resulting in insufficient precision for the early identification of POD. In recent years, the introduction of multimodal physiological data, such as metabolomics, has provided new insights for achieving higher-precision POD prediction.

The Unique Value of Omics Technologies in Identifying POD

Recent advances in multi-omics technologies—including genomics, transcriptomics, proteomics, metabolomics, and lipidomics—have opened new avenues for unbiased and comprehensive research of pathophysiological mechanisms underlying delirium [29]. By comparing biological fluid or tissue samples from patients with POD and those without POD, these methods help identify novel biomarkers, biological pathways, and molecular features that traditional clinical indicators may fail to capture [29]. Integrated multi-omics analyses combining ML and systems biology approaches hold particular promise for elucidating the complex, multifactorial nature of POD and are crucial for developing predictive models with high precision. In this review, the application and potential of omics technologies in POD prediction will be illustrated using metabolomics as an example.

The Unique Value of Metabolomics in Identifying Preoperative Metabolic Vulnerability

Surgical trauma can trigger intense catabolic stress, resulting in widespread metabolic disturbances. However, it is challenging to comprehensively identify these metabolic disturbances through examinations in routine practice. By high-throughput monitoring of hundreds of small-molecule metabolites in blood or urine, metabolomics can systematically capture preoperative nutritional status, inflammation, and metabolic dysregulation signals that emerge in the early postoperative phase. For example, aberrant tryptophan-kynurenine metabolism is associated with neuroinflammation and cognitive impairment [30]. In a recent prospective multicenter study, 260 older adults undergoing hemiarthroplasty of the hip were recruited, with samples collected; therefore, a prediction model for POD was subsequently developed by integrating high-throughput targeted metabolomics based on ML. Preoperative serum samples were analyzed using high-throughput targeted metabolomics to identify differentially expressed metabolites. Subsequently, variables were further screened using random forest and LASSO regression. Based on the screened variables, multiple methods (gradient boosting, random forest, and logistic regression) were employed to construct predictive models. The study demonstrated that the model built with logistic regression performed best, with an AUC ranging from 0.844 to 0.856 [31]. Therefore, incorporating preoperative metabolomics and ML can achieve more accurate prediction of POD in older adults with hip fractures, facilitating personalized risk stratification and clinical management.

Although metabolomics and other omics approaches hold great promise, there are currently no fully validated biomarkers available for use in routine practice for risk stratification of POD [32]. The complexity of pathophysiological mechanisms of POD may be one of the possible reasons. The mechanisms involve neuroinflammation, blood-brain barrier dysfunction, neurotransmitter imbalances, and neurodegenerative changes; moreover, these mechanisms interact with one another [33]. Therefore, using a single biomarker may be challenging to comprehensively reflect the risk of delirium. Another possible culprit may stem from the substantial cohort heterogeneity (varying by surgery type, population, and methods for delirium assessment) and small sample sizes, leading to inconsistent results [33]. Moreover, most newly discovered biomarkers require independent external validation in different clinical settings prior to translation into clinical practice [33]. These challenges highlight the need for larger-scale, more standardized multicenter studies, and systematic approaches to biomarker discovery and validation.

In clinical practice, leveraging a minimal set of core biomarkers can facilitate rapid, cost-effective, and reproducible testing, and this has been attempted in several real-world studies. A prospective study involving 172 older adults undergoing laparoscopic surgery found that preoperative plasma p-tau231 had an AUC of 0.966 for predicting POD, with a sensitivity of 0.900 and a specificity of 0.967, outperforming other tau protein subtypes (p-tau181, p-tau217, and T-tau) [34]. Similarly, a metabolomics study utilizing genetic algorithms and artificial neural networks identified L-lactate, inositol, and methylcysteine as the most salient primary metabolic biomarkers, achieving an AUC of 0.99 in predicting POD [35]. These findings suggest that a combination of two to three core biomarkers may be sufficient to provide the predictive capability required for clinical applications, while substantially reducing costs, shortening testing time, and minimizing logistical barriers compared to comprehensive omics analyses. Again, these encouraging findings need to be further validated in larger, multicenter cohorts prior to real-world applications in routine practice.

Beyond validation, once a reliable set of core biomarkers has been identified, future research must prioritize the standardization of pre-analytical processing, the control of interlaboratory batch effects, and the development of rapid point-of-care testing platforms to ensure clinically meaningful turnaround times for reporting (such as within 1–2 hours before surgery). Unless these practical limitations are addressed, even the most accurate metabolomics models will be unable to inform perioperative decision-making.

Integrating Preoperative and Intraoperative Data Further Improves Predictive Accuracy

A recent study further emphasized that the development of POD may result from the combined effects of preoperative and intraoperative factors. In this prospective observational study involving 536 patients aged ≥70 years, preoperative characteristics and intraoperative acute stressors were integrated to construct a predictive model, including demographics, frailty assessment, cognitive function, medication counts, laboratory test results, surgical invasiveness, duration of surgery and anesthesia, intraoperative fluid management, and immediate postoperative medication use. Regarding delirium-related variables, the study accounted for the multidimensional nature of delirium by collecting data on its onset, subtype, duration, severity, and time of onset, thereby facilitating comprehensive prediction of POD. A predictive model was constructed based on a multi-output LSTM model based on the collected data. The results indicated that the comprehensive inclusion of intraoperative and preoperative data substantially improved the prediction of delirium subtypes, duration, and onset time. The results indicated that incorporating intraoperative data alongside preoperative indicators substantially improved predictive performance. Specifically, the weighted-average AUC (across all five delirium-related outputs) increased from 0.76 to 0.81 [36]. This study offers new insights for enhancing risk stratification.

The Clinical Value of Emerging Digital Tools for Assessing POD Risk

The Shift From Macro-Level Profiling to Micro-Level Risk Quantification

The risk of POD is determined by the individual's capacity to tolerate or adapt to the stress imposed on the central nervous system by surgical trauma and anesthetic agents (that is, the risk of POD is directly determined by an individual's underlying resilience to stress and their vulnerability). Multiple factors such as age-related cognitive decline, neurodegenerative changes, and increased burden of comorbidities collectively render the brain less adaptive to stress, thereby enhancing the risk of functional impairment transitioning to clinical delirium among older adults when exposed to external stressors.

However, in the aforementioned ML models, the inputs are typically limited to clinical variables collected at a single time point (e.g., age, comorbidities, medication history, and ASA classification). Accordingly, the model's outputs provide only a static snapshot of the patient's overall disease burden and physiological status (i.e., a list of the patient's diseases and a rough indication of organ functions). However, this macro-level profile, derived from routinely available clinical variables such as age, ASA classification, and comorbidities, does not directly capture individual resilience and quantify the neurophysiological reserve required by older patients for coping with stress.

Although indicators of vulnerability for stress tolerance provide incremental predictive value beyond standard clinically available variables, they are often not derived from routine clinical data. For example, frailty typically requires additional assessment. Moreover, systematic reviews indicate that frailty is an independent risk factor for POD in older adults [37,38]. Frailty is defined as a state of increased vulnerability to stressors resulting from age-related multisystem decline, involving multiple domains such as musculoskeletal, metabolic, neuroendocrine, and immune systems [39]. The gold standard for frailty assessment is the Comprehensive Geriatric Assessment (CGA)—a multidimensional, multidisciplinary evaluation instrument that covers physical health, functional status, mental health, and social circumstances [39]. However, the CGA is resource-intensive and typically takes 45–90 minutes to complete, posing a considerable barrier to implementation in fast-paced surgical settings [39]. Consequently, brief screening tools are commonly adopted in clinical practice, such as the Fried Frailty Phenotype (approximately 20–30 minutes) or the FRAIL scale (approximately 5 minutes); however, this is accompanied by reduced comprehensiveness of frailty assessment [39].

Similarly, cognitive reserve—which reflects the brain's capacity to tolerate pathological damage and stress to maintain homeostasis—is closely associated with POD risk [40]. Cognitive reserve refers to the brain's ability to cope with or compensate for age-related neurobiological changes or pathological damage, thereby maintaining cognitive function in the presence of underlying neuropathology [41]. Cognitive reserve is often indirectly assessed by proxy measures that reflect lifelong cognitive experiences, including educational attainment, occupational complexity, engagement in leisure activities, and social participation [42]. The Cognitive Reserve Index Questionnaire (CRIq) is one of the most widely used tools. However, CRIq assessments are time-consuming due to the lengthy instrument used and require trained assessors [43].

Moreover, a recent review indicates that delirium is associated with disruptions in daily activity rhythms, including reduced daytime activity, increased nighttime activity, and fragmentation of the sleep-wake cycle [44]. After incorporating motor characteristics, an ML-based model improved its predictive accuracy from approximately 62% to 74%, indicating that motor patterns can provide additional information beyond the routine clinical variables [44]. Therefore, to further improve the predictive accuracy of existing models, it is essential to incorporate the vulnerability indicators into risk stratification.

However, in clinical practice, assessments are often conducted by trained professionals based on specialized scales (for frailty and cognitive function) or rely on fragmented information derived from the recollections of patients or caregivers (regarding daily activity patterns). Additionally, such assessments remain static and snapshot-like, failing to reflect patients' actual functional status and fluctuation patterns in daily life. Subjective reports are also influenced by recall bias and social expectations.

Emerging digital tools have the potential to transform the collection and assessment of these vulnerability indicators. AI agents and chatbots can automatically administer standardized cognitive assessments through voice-based conversations, transforming cognitive screening into an automated process that can be integrated into routine practices. As a result, comprehensive preoperative cognitive function assessments can be implemented with minimal clinical burden. By recording step counts, activity segments, and sleep-wake rhythms over several consecutive days, wearable sensors transform frailty and daily activity capacity from a single subjective impression into objective, continuous, and quantifiable dynamic trajectories—information that conventional assessments cannot provide. EEG monitoring can non-invasively capture subtle changes in resting-state cortical electrical activity, identifying latent neurophysiological vulnerabilities before detectable declines in macro-level cognitive function occur.

A valid concern is that the ever-increasing stream of continuous, dynamic data from digital monitoring tools may complicate clinical decision-making for critically ill or unstable patients. Indeed, unprocessed raw data streams may lead to information overload and alarm fatigue, increasing the risk of delayed interventions [45]. However, algorithms built into digital health tools can integrate raw multimodal data, filter out noise, and ultimately present actionable, decision-support insights—thereby reducing the cognitive burden. This advantage is particularly important in the context of POD, a condition whose symptoms fluctuate dramatically within a matter of hours. Recent studies have shown that, when using digital tools, continuous physiological and activity monitoring can effectively capture early signs of delirium, achieving robust predictive performance that outperforms static, intermittent assessments [44,46]. Therefore, the real challenge lies not in the volume of data itself, but in how the data is presented and integrated into clinical workflows—ensuring that digital tools enhance, rather than overwhelm, clinical judgment.

Large Language Models and Chatbots Can Automate Standardized Cognitive Screening

Since the breakthroughs in large language models (LLMs) at the end of 2022, a transformation is underway: from human-work interactions to human-agent collaboration. Thanks to their excellent natural language understanding and generation capabilities, as well as their vast knowledge bases, LLMs can perform tasks that previously required specially trained humans. Therefore, LLMs hold the potential to automate certain components of daily practice, thereby improving efficiency, reducing healthcare costs, and increasing access to care.

Chatbots hold promise for automating preoperative assessment of cognitive function in older adults by leveraging standardized measurement tools. The Telephone Interview for Cognitive Status-Modified (TICS-M) is a widely used remote cognitive assessment tool. A recent study investigated the effectiveness of conversational AI agents in cognitive function assessment (based on the TICS-M tool) by comparing the psychometric properties of the tool when administered by psychologists or conversational AI agents. The results indicated that the psychometric properties of the TICS-M as assessed by conversational AI agents are comparable to those assessed by psychologists (including concurrent validity and test-retest reliability). Therefore, in the context of preoperative assessments, AI can be used to automate cognitive function evaluations, as the results are reliable, valid, and safe, while offering the advantages of low cost and greater accessibility [47]. It is worth noting that the rapid evolution of mainstream LLMs in human-computer interaction has made automated cognitive function assessment that utilizes LLMs in older adults increasingly feasible. Many models support natural language interaction with users and generate visualizations to aid in conveying information. In this context, older adults are no longer required to master complex operative commands; instead, they can complete cognitive function assessments through conversational interactions, which substantially reduce the cognitive barrier and anxiety associated with using digital tools.

Another study developed an agent for delirium screening tailored for the ICU setting. The researchers first constructed a knowledge graph specific to delirium based on published clinical guidelines, and then built an agent capable of simulating clinical decision-making pathways to generate outputs, including selection of measurement tools, confidence levels for conclusions, and guideline-based recommendations for delirium care. Performance testing of the agent was conducted on 20 patients, with results showing that the agent could select appropriate tools with high accuracy and achieved an overall accuracy of 90% in generating conclusions and care recommendations, whereas other LLMs achieved only 45% [48]. Therefore, this agent holds promise for reducing missed or incorrect diagnoses, and its greatest advantage lies in integrating assessment tools and relevant guidelines into daily practice [49].

A Comparative Analysis of the Performance of General-Purpose Large Language Models and Domain-Specific Models in Perioperative Decision-Making

The study of this delirium screening agent development also highlights a substantial performance gap in accuracy between domain-optimized agents and general-purpose LLMs (90% vs. 45%) [48]. This disparity stems mainly from the domain-optimized agent's incorporation of structured domain knowledge graphs and clinical guideline rules, which constrain its outputs to an evidence-based framework relevant to delirium, rather than statistical correlations learned from massive corpora (which can easily misclassify unrelated elements as related). This comparison highlights a key principle for applying AI to perioperative management: when clinical decisions require extreme precision and work within tight time windows, the specialized expertise of domain-optimized models proves more reliable. Delirium screening—particularly the identification of the hypoactive type—is a classic scenario requiring both high accuracy and strong timeliness: clinicians must make decisions within minutes based on fragmented information; in such conditions, an AI agent pre-trained on delirium-specific guidelines is more likely to provide accurate, compliant, and timely decision support.

However, even in domain-specific models, reliability remains far from perfect. A recent study compared the performance of six LLMs in providing decision support for perioperative care of older adults with multiple comorbidities. These six models included five general-purpose LLMs (ChatGPT, Gemini, DeepSeek, Claude, Kimi) and one domain-optimized model (New Youth Anesthesia AI Assistant, NYAAI). The (simulated) case involved an 84-year-old male with multiple comorbidities and a femoral fracture. Although NYAAI showed the best overall performance, it revealed only partial compliance with guidelines (specifically overlooking temperature management and delirium); the general-purpose LLMs, meanwhile, generated inappropriate recommendations [50]. Therefore, based on current evidence, while LLMs can automate standardized assessments and generate structured recommendations for addressing coverage and consistency gaps in traditional manual evaluations, their results still require verification by healthcare professionals to ensure the accuracy of clinical decisions. The broader challenges and risks associated with the clinical application of LLMs, including hallucinations, interpretability, privacy, and medical liability, have been thoroughly discussed elsewhere [51].

Wearable Sensors Enable Continuous, Objective Monitoring

Behavioral Patterns and Frailty Trajectories

Since preoperative declines in physical activity and frailty are risk factors for POD, the use of wearable sensors to continuously monitor physical activity and frailty status for indirect POD risk stratification represents an important aspect of digital technology in the context of POD. Surgical trauma itself acts as a stressor, leading to more persistent metabolic disturbances in frail older adults; furthermore, recovery of the central nervous system is highly dependent on energy and the immune microenvironment, rendering preoperative frailty an independent risk factor for POD. Physical activity performance can serve as a measurable, observable indicator of frailty. Walking speed can be used to assess frailty [52]. Furthermore, a decline in physical activity performance reflects, to some extent, degenerative changes in the neuromuscular control system.

In routine practice, assessments of frailty and physical activity capacity typically rely heavily on subjective judgment or patient self-reports rather than a comprehensive understanding of fluctuations in the patient's daily condition, and such evaluations are likely less accurate because a single measurement may not be able to fully capture the actual state of these two indicators. In particular, self-reports are prone to recall bias and social expectations, and patients with preoperative cognitive impairment may be unable to provide accurate reports. Wearable sensors can further enhance the reliability of clinical decision-making by continuously and objectively recording daily activity, walking speed, heart rate variability, and movement patterns, thereby enabling more precise and comprehensive preoperative risk identification than conventional assessments alone.

A prospective observational study, using a preoperative clinical assessment scale and the 6-minute walk test as reference standards, evaluated for the first time whether data generated by wearable devices could be used for preoperative risk assessment. All participants (older adults scheduled to undergo a 2-hour surgery under general anesthesia) were required to wear wearable devices continuously for 7 days before surgery. Findings indicated that daily step count and VO_2max recorded by wearable devices were significantly correlated with the performance of the 6-minute walk test and the clinical assessment scale [53]. Another cross-sectional study used data generated by a thigh-worn accelerometer worn continuously for 10 days to assess whether older adults were in a pre-frail or frail state, with the Fried Frailty Phenotype (FFP) and the Comprehensive Geriatric Assessment-based Frailty Index (CGA-FI) serving as reference criteria. The findings indicate that such wearable devices can effectively identify pre-frail or frail states and hold promise as a new frailty screening tool [54]. A study based on UK Biobank data evaluated digital gait biomarkers generated by wrist-worn sensor devices in middle-aged and older adults and found that they serve as an effective and simple tool for frailty assessment, with performance comparable to that of the FFP [55].

It is important to note that the current evidence primarily supports the use of activity metrics obtained from wearable devices as indicators of frailty and functional vulnerability, rather than as direct predictors of POD. The logical pathway is essentially indirect: wearable sensors collect continuous, objective data on physical activity, gait, and circadian rhythms—information that reflects an individual's overall physiological resilience in the face of stress. This vulnerability, in turn, is a recognized upstream determinant of POD risk. Therefore, no digital biomarkers derived from wearable devices can directly predict POD with sufficient accuracy; however, these metrics offer unique incremental value that traditional point-in-time assessments cannot provide: they transform subjective, recall-dependent, and static frailty assessments into objective, longitudinal, and dynamic functional trajectories.

This distinction is of great importance for clinical integration. Wearable devices are not intended to serve as standalone predictors of POD, but rather as tools for continuous vulnerability monitoring, feeding objective functional data into a multimodal risk stratification framework. In this complementary role, wearable devices do not replace clinical judgment but rather extend the temporal and objective dimensions of frailty assessment beyond the capabilities of traditional tools.

Promising Indicators for POD Prediction by Wearable Devices

The association between heart rate variability and postoperative delirium in older adults undergoing cardiovascular surgery and admitted to the ICU was systematically confirmed, indicating that heart rate variability, as an important parameter of autonomic nervous system activity, is associated with POD development and is an independent risk factor for predicting adverse postoperative outcomes [56]. Another study further displayed the capability of wearable devices in the continuous monitoring of heart rate variability in routine practice, demonstrating that wearable devices can effectively capture the diurnal dynamic changes of heart rate variability [57].

Also, given that both mild cognitive impairment (MCI) and POD involve alterations in cognitive function, wearable EEG devices—which have been validated for detecting MCI—may also offer a promising option for the high-precision prediction of POD [58], as the relationship between certain EEG features (such as the reduction in alpha wave power in the occipital lobe) and POD has been confirmed [59].

Moreover, sleep loss and sleep disturbances are also associated with POD [60,61]. Wearable devices can estimate sleep by using accelerometers and heart rate sensors, and can be used in sleep monitoring for patients after head and neck surgery [62]. The use of wearable devices in the ICU setting has been reviewed in detail elsewhere [63].

EEG Monitoring Enables Preoperative Identification of Latent Vulnerability and Intraoperative Deep Regulation

EEG patterns reflect neural susceptibility at the cortical electrophysiological level. Cortical electrical activity recorded via EEG is highly sensitive to insufficient cerebral perfusion, metabolic disturbances, and neurotransmitter imbalances—all of which are central components of the underlying pathophysiology of POD. In older adults, resting-state EEG exhibits specific changes with aging (decreased α-band power and increased slow-wave activity in the θ and δ bands), reflecting reduced cortical activation due to cholinergic neuron degeneration. The perioperative EEG features that can be used to predict POD have been described in detail elsewhere [59,64,65]. ML-based algorithms can automatically extract and quantify relevant EEG patterns, thereby identifying preoperative brain changes that are clinically subtle at the neurophysiological level and providing information on neural reserve that cannot be obtained from conventional clinical variables.

A small-scale prospective observational cohort study preliminarily explored the feasibility of using preoperative resting-state EEG and MoCA assessment scores, which reflect the overall functional performance of the brain, as predictors of POD in older adults, thereby facilitating targeted interventions. The study indicated that occipital α-band power and MoCA-based cognitive assessment were effective predictors of POD (AUC = 0.94, 95% CI [0.82–0.98]), with a sensitivity of 0.83 and a specificity of 0.91. This study further confirmed that the thalamocortical circuit exhibits distinct EEG patterns under stressors, validating the feasibility of high-precision prediction using the EEG [66]. Given the challenges of using standard 20-lead EEG for screening in routine practice, one study developed an ML-based POD prediction model based on preoperative data from portable EEG and cognitive function assessments, achieving an accuracy of 0.86 and an AUC of 0.93. The finding of this study further supports the feasibility of generating highly accurate predictions without the need for expensive equipment [67].

The key value of intraoperative EEG monitoring lies in its ability to capture the brain's neurophysiological response to surgical stress under anesthesia in real time, thereby providing an objective risk stratification for the identification of POD. The European Society of Anaesthesiology and Intensive Care Medicine (ESAICM) guidelines recommend that anesthesiologists can independently interpret raw intraoperative EEG traces [68]. Specifically, intraoperative burst-suppression—a sign of excessive anesthetic depth, characterized by alternation between high-amplitude burst activity and low-voltage suppressive activity on the EEG—is independently associated with a 41% increase in the relative risk of POD [69]. Based on this association, the ESAICM guidelines recommend monitoring this EEG pattern to guide anesthetic titration [68]; however, this should be viewed as a risk-relevant intraoperative warning sign rather than a diagnostic or identification approach for POD itself.

A retrospective study explored the feasibility of combining intraoperative EEG and clinically available data from routine practice to predict POD in older adults based on an ML-based algorithm. In this study, balanced ensemble methods were used to address data imbalance, and a predictive model was constructed. Furthermore, participants were grouped based on the type of anesthetic used. The study found that adding EEG-related features improved model performance: the AUC for clinical data alone was 0.75, while in the model integrating intraoperative EEG (especially in participants receiving sevoflurane to attain anesthetic effects), the AUC reached as high as 0.80 [70].

Structural Challenges and Future Prospects for Digital Technologies in Perioperative Management

Although digital tools, including ML-based risk prediction models, AI agents, and wearable devices, have shown promise in identifying POD risks preoperatively and intraoperatively, translating these advances into clinical decision-making tools for routine practice faces a variety of structural challenges.

Sample Size Challenges and Data Bias Risks

Machine learning and even deep learning algorithms learn the intrinsic relationships between diseases and predictive factors through a data-driven approach, which empowers them to handle the complex, nonlinear interactions among EEG patterns, metabolite alterations, and clinical variables. However, this nature raises a key concern: the patterns learned by the model depend entirely on the quality and representativeness of the input data. Currently, multimodal predictive models in the field of POD risk identification, including those integrating EEG or metabolomics data, are generally constructed based on small sample sizes. With such sample sizes, ML models are highly prone to instability and overfitting, and most studies have failed to clarify the theoretical basis for sample size calculations. This methodological gap presents more challenges for assessing the reliability boundaries of current models.

Furthermore, a larger dataset volume does not necessarily indicate superior data quality; rather, data quality must be assessed in the context of the model's intended condition. For example, data from the UK Biobank is widely recognized as high-quality. However, participant recruitment for this database is based on voluntary registration, which may introduce a healthy volunteer bias: participants tend to be healthier and have higher socioeconomic status than the general population, potentially leading to an overestimation of overall functional status among older adults. Therefore, it is reasonable to infer that the population included in this database differs from older adults in terms of comorbidity burden, functional status, and postoperative stress response. Within this context, even if future studies leverage UK Biobank data to develop a wearable device–based digital biomarker model for predicting POD in older adults, this model would still require further validation in local hospitals to determine the optimal target population and reliability boundaries of this digital tool.

Challenges Posed by the Heterogeneity of Multi-Source Data and Standardization

Multimodal fusion, marked by the integration of EHRs, EEGs, metabolomics, and wearable device data into a single framework, faces practical obstacles due to inconsistent data formats and standards. Other challenges include: (i) different hospitals may report the same laboratory test using different measurement units; (ii) EEG devices from different manufacturers may have varying lead configurations; and (iii) wearable device algorithms may frequently change across software updates. The heterogeneity of multi-source data may introduce unexpected confounding factors, leading to poor performance of digital tools when deployed across different settings. Therefore, it will be necessary in the future to establish unified data standards and data cleaning processes to clarify how multi-source data are integrated and standardized in the study, and to simultaneously address data quality, representativeness, and bias. Furthermore, since medical data often contains errors, missing values, or inconsistencies, few current AI models explicitly describe the distribution and types of missing values in their methodology, whether data imputation methods were used, and their potential impact on results.

Feasibility, Practical Constraints, and Strategy Selection for a General-Purpose POD Prediction Model Across Surgical Types

Currently, the majority of POD prediction models have been developed and validated for a particular surgical type or closely related surgery types (such as orthopedic surgery alone, cardiac surgery alone, or non-cardiac surgery as a broad category). Given the wide variety of surgeries performed on older adults, a reasonable and compelling question arises: Is it possible to develop a general-purpose POD prediction model that covers multiple types of surgeries?

Theoretically, this goal is not unattainable. By incorporating surgical procedure types and their associated characteristics, such as the degree of surgical trauma, expected duration, surgical site, and intraoperative blood loss, as model inputs, the algorithm can learn the relative contributions of different surgical procedures to the risk of POD, thereby enabling risk stratification across surgical procedures within an integrated framework. In theory, if the model is exposed to sufficiently diverse data across surgical types during the training phase, it may possess stronger generalization capabilities and a broader clinical applicability than models designed for a single surgical type.

First, the required sample size is extremely large. In the development of clinical predictive models, it is typically required that each predictor variable correspond to at least 10–20 outcome events (events per variable, EPV). To cover 5–6 types of surgery in a single model and incorporate surgery-specific features and interaction terms, the total sample size required may reach thousands or even tens of thousands of cases, with a particularly high demand for POD-positive cases. However, most current POD prediction studies in this field have sample sizes of only a few hundred cases, representing an order-of-magnitude gap from the requirements for building a stable, general-purpose model.

Second, there are fundamental differences in the pathophysiological mechanisms underlying POD across different types of surgery, posing a severe challenge to the universal applicability of the model. POD following cardiac surgery may be closely associated with microembolism and systemic inflammatory responses induced by cardiopulmonary bypass; major orthopedic surgeries involve factors such as fat embolism, postoperative immobilization, and pain; and major abdominal surgeries are accompanied by severe systemic inflammatory responses and metabolic disturbances. These mechanistic differences imply that the weightings of the same set of predictors may vary significantly across different surgical procedures. Whether a "universal model" can maintain sufficient predictive accuracy under such complex and heterogeneous conditions is, by itself, a scientific question that requires rigorous empirical testing—not one that can be resolved simply by stacking technologies.

Third, there are considerable obstacles to obtaining and integrating data across different surgical procedures. Data on different surgical procedures are typically scattered across different surgical departments or centers, and there may be variations in data collection standards, variable definitions, delirium assessment tools, and assessment frequency. Effectively integrating such heterogeneous data requires addressing a series of methodological challenges, including data standardization, handling of missing values, and control of confounding factors.

Fourth, research resources and funding are crucial aspects. Investigating multiple types of surgery simultaneously in a prospective multicenter study typically entails heightened complexity and costs related to study coordination, data collection, quality control, and ethical review, creating a burden that is prohibitive for the vast majority of research teams.

Based on the above considerations, we recommend a phased approach: at this stage, prioritize the development, internal validation, and prospective external validation of models within a single surgical procedure or among populations with similar procedures to ensure their reliability and clinical utility; once sufficient evidence has been accumulated for models specific to a single surgical procedure, gradually explore the optimal architecture and feasible boundaries of a universal model across surgical procedures through large-scale, multicenter collaborative studies. This strategy not only enables a pragmatic approach to research but also leaves room for future integration.

The Black-Box Dilemma of Algorithms and Trust in Clinical Human-Machine Interaction

Given the substantial variations in clinical interventions resulting from different POD risk thresholds, such models should be highly interpretable. The outputs of digital tools for POD risk identification directly influence decisions regarding whether to postpone surgery for preoperative optimization, shift from general to regional anesthesia, or initiate costly multidisciplinary and rehabilitation interventions. Although some AI experts are currently working on enhancing algorithmic interpretability (such as Shapley Additive explanation (SHAP) values for feature contribution decomposition), deep learning/ML-based digital risk stratification tools remain challenging to integrate into routine practice due to the black-box nature of AI. Moreover, the model's outputs are not always accurate and still require correction and fine-tuning based on clinicians' expertise. Limited interpretability is particularly problematic and concerning when algorithmic outputs conflict with clinical intuition. In this situation, clinicians may overlook actual high-risk patients due to distrust, or blindly follow the model's misjudgments due to over-reliance.

To address this, future research should explore variations in clinicians' decision-making behaviors when faced with outputs from different explainability tools (SHAP, LIME, rule extraction), and evaluate the impact on clinical acceptance and decision accuracy when they are exposed to the combined presentation of digital tool outputs and feature explanations, thereby determining the optimal human-machine interaction approach. At present, the most appropriate workflow entails digital tool–driven risk evaluations and explanations, as well as final decision-making by clinicians following the integration of information obtained from multiple sources.

To further promote the integration of digital tools into routine practice, future research should validate the combination of preoperative risk prediction and specific interventions, and evaluate the clinical outcomes and cost-effectiveness of the bundled strategy (risk prediction and treatment) in randomized controlled trials.

Conclusions

Postoperative delirium (POD) remains a major challenge in geriatric surgical care, yet it is often underdiagnosed. Digital technologies are playing an increasingly prominent role in identifying perioperative risk of POD in older adults. These digital approaches can effectively automate risk stratification, capture neurophysiological and functional vulnerabilities that traditional assessments cannot detect, and provide objective, continuous data inputs for multimodal risk frameworks. However, it is important to recognize that existing research primarily supports their utility in risk stratification and quantification rather than as validated interventions capable of improving clinical outcomes. To date, there is no evidence that these digital tools alone can directly reduce the incidence of postoperative complications, such as POD, or improve long-term patient outcomes.

Several methodological challenges must be addressed before clinical translation can be achieved. Future research should prioritize prospective, multicenter validation studies to assess whether the combination of digitally assisted risk stratification and targeted preventive interventions can effectively improve patient-centered outcomes. At present, digital technologies should be viewed as promising decision-support tools with enhanced risk identification capabilities rather than as validated preventive strategies. Adopting this cautious and forward-looking perspective will facilitate the responsible integration of digital tools into perioperative care for older adults.

Availability of Data and Materials

Not applicable.

Author Contributions

YL, MLZ, and WYD conceived the review, designed the search strategy, and drafted the manuscript. TML, CLL, XYW, and HYJ required data, analyzed data, performed the formal analysis, designed the methodology, reviewed successive drafts, and provided methodological guidance and editorial feedback. All authors have been involved in revising the manuscript critically for important intellectual content. All authors gave final approval of the version to be published. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

We are gratefully acknowledge the support and assistance provided by the Zhejiang Provincial Medical and Health Technology Project (Grant No. 2024KY1304) and The Construction Fund of Key Medical Disciplines of Hangzhou (Grant No. 2025HZZD05).

Funding

This research was supported by Medical and Health Technology Project of Hangzhou (grant. A2026339).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Canales C, Wann L, Blitz J, Whittington R. The older adult surgical patient: a review of optimization and gaps in clinical practice. Perioperative Medicine (London, England). 2025; 14: 104. https://doi.org/10.1186/s13741-025-00593-x

[2] Sheng J, Jiang Q, Xie D, Li X, Chen H, Zeng C, et al. Age-Related Trends in Total Joint Arthroplasty in China: A Nationwide Study. The Journal of Arthroplasty. 2025; 40: S146–S152. https://doi.org/10.1016/j.arth.2025.08.056

[3] Baldo A, Chatzipanagiotou OP, Woldesenbet S, Akabane M, Spolverato G, Pawlik TM. New-Onset Geriatric Syndromes among Patients Undergoing Major Operation: Impact on Clinical Outcomes and Quality of Life. Journal of the American College of Surgeons. 2026; 242: 663–674. https://doi.org/10.1097/XCS.0000000000001661

[4] Smirnov A, Semionov M, Yasinski V, Binyamin Y, Zlotnik A, Frank D. Post-Operative Delirium in Elderly Patients: A Narrative Review. International Journal of Molecular Sciences. 2025; 26: 11314. https://doi.org/10.3390/ijms262311314

[5] Jin Z, Hu J, Ma D. Postoperative delirium: perioperative assessment, risk reduction, and management. British Journal of Anaesthesia. 2020; 125: 492–504. https://doi.org/10.1016/j.bja.2020.06.063

[6] Lv S, Li J, He H, Zhao Q, Jiang Y. Artificial intelligence applications in delirium prediction, diagnosis, and management: a systematic review. Artificial Intelligence Review. 2025; 58: 269. https://doi.org/10.1007/s10462-025-11264-0

[7] Tesh RA, Sun H, Jing J, Westmeijer M, Neelagiri A, Rajan S, et al. VE-CAM-S: Visual EEG-Based Grading of Delirium Severity and Associations With Clinical Outcomes. Critical Care Explorations. 2022; 4: e0611. https://doi.org/10.1097/CCE.0000000000000611

[8] Zugni N, Guadrini L, Rasulo F. Noninvasive neuromonitoring in the operating room and its role in the prevention of delirium. Best Practice & Research. Clinical Anaesthesiology. 2021; 35: 191–206. https://doi.org/10.1016/j.bpa.2020.09.006

[9] Kunicki ZJ, Ngo LH, Marcantonio ER, Tommet D, Feng Y, Fong TG, et al. Six-Year Cognitive Trajectory in Older Adults Following Major Surgery and Delirium. JAMA Internal Medicine. 2023; 183: 442–450. https://doi.org/10.1001/jamainternmed.2023.0144

[10] Surgical Nursing Professional Committee of Chinese Nursing Association; Nursing Department/Neurosurgery/Vascular Surgery/Critical Care Medicine, The First Affiliated Hospital of Chongqing Medical University; Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. Expert consensus on the prevention and nursing of postoperative delirium in adults. Chinese Journal of Emergency and Critical Care Nursing. 2025; 6: 837–842. https://jwzz.zhhlzzs.com/EN/10.3761/j.issn.2096-7446.2025.07.011

[11] Lin CJ, Fick DM, Traynor V, Chen YC, Hsiang HF, Chiu HY. Comparative Diagnostic Accuracy of Nursing Delirium Screening Scale Versus Confusion Assessment Method for Postoperative Delirium: A Systematic Review and Meta-Analysis. Journal of Clinical Nursing. 2025; 34: 287–298. https://doi.org/10.1111/jocn.17467

[12] Maybrier HR, Mickle AM, Escallier KE, Lin N, Schmitt EM, Upadhyayula RT, et al. Reliability and accuracy of delirium assessments among investigators at multiple international centres. BMJ Open. 2018; 8: e023137. https://doi.org/10.1136/bmjopen-2018-023137

[13] Ahmadzadeh S, Duplechin DP, Haynes AT, Hollander AV, Rieger R, Jean Baptiste C, et al. Evolving Clinical Management of Postoperative Delirium: A Narrative Review. Cureus. 2025; 17: e92927. https://doi.org/10.7759/cureus.92927

[14] Kong W, Jiang J, Wang Y, Chen J, Wang B, Wang K, et al. Machine learning-based predictive model for postoperative delirium of elderly patients with coronary heart disease undergoing non-cardiac surgery: a retrospective cohort study. Frontiers in Psychiatry. 2026; 17: 1780056. https://doi.org/10.3389/fpsyt.2026.1780056

[15] Matsumoto K, Nohara Y, Sakaguchi M, Takayama Y, Fukushige S, Soejima H, et al. Temporal Generalizability of Machine Learning Models for Predicting Postoperative Delirium Using Electronic Health Record Data: Model Development and Validation Study. JMIR Perioperative Medicine. 2023; 6: e50895. https://doi.org/10.2196/50895

[16] Song YX, Yang XD, Luo YG, Ouyang CL, Yu Y, Ma YL, et al. Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: A retrospective study. CNS Neuroscience & Therapeutics. 2023; 29: 158–167. https://doi.org/10.1111/cns.13991

[17] Silvey S, Liu J. Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study. Journal of Medical Internet Research. 2024; 26: e60231. https://doi.org/10.2196/60231

[18] Tsegaye B, Snell KIE, Archer L, Kirtley S, Riley RD, Sperrin M, et al. Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: methodological systematic review. Journal of Clinical Epidemiology. 2025; 180: 111675. https://doi.org/10.1016/j.jclinepi.2025.111675

[19] Hu Y, Zhang X, Slavin V, Belsti Y, Tiruneh SA, Callander E, et al. Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality. Journal of Medical Internet Research. 2025; 27: e77721. https://doi.org/10.2196/77721

[20] Ma B, Shi C. Beyond AUC: Questioning the clinical interpretation of “superior” discrimination in cardiovascular health scores. International Journal of Cardiology. Cardiovascular Risk and Prevention. 2026; 29: 200592. https://doi.org/10.1016/j.ijcrp.2026.200592

[21] Benitez-Aurioles J, Wynants L, Peek N, Goodley P, Crosbie P, Sperrin M. The continuous net benefit: assessing the clinical utility of prediction models when informing a continuum of decisions. Diagnostic and Prognostic Research. 2026; 10: 8. https://doi.org/10.1186/s41512-026-00224-z

[22] van Leeuwen FD, Steyerberg EW, van Klaveren D, Wessler B, Kent DM, van Zwet EW. Instability of the AUROC of Clinical Prediction Models. Statistics in Medicine. 2025; 44: e70011. https://doi.org/10.1002/sim.70011

[23] Feng G, Xu H, Wan S, Wang H, Chen X, Magari R, et al. Twelve practical recommendations for developing and applying clinical predictive models. The Innovation Medicine. 2024; 2: 100105. https://doi.org/10.59717/j.xinn-med.2024.100105

[24] Kocar TD, Wolf P, Leinert C, Brefka S, Fotteler ML, Uihlein A, et al. SURGE-ahead postoperative delirium prediction: external validation and open-source library. European Geriatric Medicine. 2025; 16: 851–859. https://doi.org/10.1007/s41999-025-01180-5

[25] Holler E, Ludema C, Ben Miled Z, Rosenberg M, Kalbaugh C, Boustani M, et al. Development and Validation of a Routine Electronic Health Record-Based Delirium Prediction Model for Surgical Patients Without Dementia: Retrospective Case-Control Study. JMIR Perioperative Medicine. 2025; 8: e59422. https://doi.org/10.2196/59422

[26] Liu J, Liu X, Zhang C. Analysis of Influencing Factors and Construction of Prediction Model for Delirium After Intraspinal Anesthesia in Elderly Patients Undergoing Femoral Neck Fracture Surgery. Annali Italiani Di Chirurgia. 2026; 97: 317–328. https://doi.org/10.62713/aic.4257

[27] Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. Journal of Clinical Epidemiology. 2016; 79: 76–85. https://doi.org/10.1016/j.jclinepi.2016.05.007

[28] Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Medicine. 2023; 21: 70. https://doi.org/10.1186/s12916-023-02779-w

[29] Vasunilashorn SM, Dillon ST, Marcantonio ER, Libermann TA. Application of Multiple Omics to Understand Postoperative Delirium Pathophysiology in Humans. Gerontology. 2023; 69: 1369–1384. https://doi.org/10.1159/000533789

[30] Zhang Q, Zhao W, Yun Y, Ma T, An H, Fan N, et al. Multiomics analysis reveals aberrant tryptophan-kynurenine metabolism and immunity linked gut microbiota with cognitive impairment in major depressive disorder. Journal of Affective Disorders. 2025; 373: 273–283. https://doi.org/10.1016/j.jad.2024.12.070

[31] Zhong G, Huang X, Li C, Wang D, Huang D, Sun M, et al. Integrated High-Throughput Targeted Metabolomics and Machine Learning for Early Prediction and Prevention of Postoperative Delirium in Older Adult Surgical Patients: Prospective Multicenter Cohort Study. JMIR Aging. 2026; 9: e78495. https://doi.org/10.2196/78495

[32] Arai N, Sugiura Y, Nakajima S, Wada M, Moriyama S, Mimura Y, et al. Prediction of postoperative delirium by blood metabolome analysis. Journal of Psychiatric Research. 2025; 184: 500–514. https://doi.org/10.1016/j.jpsychires.2025.03.028

[33] Leng K, Maze M, Barreto Chang OL. Emerging biomarkers of postoperative delirium at the intersection of neuroinflammation and neurodegeneration. Frontiers in Aging Neuroscience. 2025; 17: 1632947. https://doi.org/10.3389/fnagi.2025.1632947

[34] Wu L, Su J, Li G, Zhuang M, Sun H, Zhang J. Preoperative Plasma p-tau231, p-tau181, and p-tau217 Are Associated With Postoperative Delirium: A Prospective Study. The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences. 2025; 80: glaf058. https://doi.org/10.1093/gerona/glaf058

[35] Ivanisenko VA, Rogachev AD, Makarova ALA, Basov NV, Gaisler EV, Kuzmicheva IN, et al. AI-Assisted Identification of Primary and Secondary Metabolomic Markers for Postoperative Delirium. International Journal of Molecular Sciences. 2024; 25: 11847. https://doi.org/10.3390/ijms252111847

[36] You J, Kim J, Choi J, Koo BN, Lee H. Multi-output LSTM-based prediction of postoperative delirium: integrating baseline and perioperative data for enhanced risk stratification in older spine surgery patients. BioData Mining. 2026; 19: 31. https://doi.org/10.1186/s13040-026-00531-7

[37] Wu H, Yan S, Cao H, Feng C, Zhang H. Unraveling the impact of frailty on postoperative delirium in elderly surgical patients: a systematic review and meta-analysis. BMC Anesthesiology. 2025; 25: 114. https://doi.org/10.1186/s12871-025-02994-3

[38] Liu CY, Gong N, Liu W. The Association Between Preoperative Frailty and Postoperative Delirium: A Systematic Review and Meta-analysis. Journal of Perianesthesia Nursing : Official Journal of the American Society of PeriAnesthesia Nurses. 2022; 37: 53–62.e1. https://doi.org/10.1016/j.jopan.2020.12.006

[39] Orkaby AR, Schwartz AW, Callahan KE. Frailty. Annals of Internal Medicine. 2026; 179: ITC17–ITC32. https://doi.org/10.7326/ANNALS-25-04412

[40] Guenther U, Hoffmann F, Dewald O, Malek R, Brimmers K, Theuerkauf N, et al. Preoperative Cognitive Impairment and Postoperative Delirium Predict Decline in Activities of Daily Living after Cardiac Surgery-A Prospective, Observational Cohort Study. Geriatrics (Basel, Switzerland). 2020; 5: 69. https://doi.org/10.3390/geriatrics5040069

[41] Stern Y, Arenaza-Urquijo EM, Bartrés-Faz D, Belleville S, Cantilon M, Chetelat G, et al. Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimer's & Dementia : the Journal of the Alzheimer's Association. 2020; 16: 1305–1311. https://doi.org/10.1016/j.jalz.2018.07.219

[42] Nucci M, Mapelli D, Mondini S. Cognitive Reserve Index questionnaire (CRIq): a new instrument for measuring cognitive reserve. Aging Clinical and Experimental Research. 2012; 24: 218–226. https://doi.org/10.3275/7800

[43] Chu ZZ, Chen XX, Qi MS, Ren JQ, Wang XJ. Cross-cultural adaptation and psychometric evaluation of the Chinese version of the short version of the Cognitive Reserve Index Questionnaire (s-CRIq): a cross-sectional study. BMC Psychiatry. 2026; 26: 431. https://doi.org/10.1186/s12888-026-08041-w

[44] Hong JS, Kim NY, Kim HR, Han DH, Kim SM. Predicting and Early Detection of Delirium through Motion Patterns: A Narrative Review. Clinical Psychopharmacology and Neuroscience : the Official Scientific Journal of the Korean College of Neuropsychopharmacology. 2026; 24: 30–39. https://doi.org/10.9758/cpn.25.1326

[45] Cecconi M, Hutanu AL, Beard J, Gonzalez-Pizarro P, Ostermann M, Batchelor A, et al. Unlocking opportunities to transform patient care: an expert insight on limitations and opportunities in patient monitoring. Intensive Care Medicine Experimental. 2025; 13: 24. https://doi.org/10.1186/s40635-025-00733-z

[46] Park C, Han C, Jang SK, Kim H, Kim S, Kang BH, et al. Development and Validation of a Machine Learning Model for Early Prediction of Delirium in Intensive Care Units Using Continuous Physiological Data: Retrospective Study. Journal of Medical Internet Research. 2025; 27: e59520. https://doi.org/10.2196/59520

[47] Serafimovska A, Swavley K, Zhang Qian Ao A, Challinor KL, Florio T. Cognitive status assessment of older adults - test administration by conversational artificial intelligence (AI) chatbot: proof-of-concept investigation. Journal of Clinical and Experimental Neuropsychology. 2025; 47: 472–484. https://doi.org/10.1080/13803395.2025.2542248

[48] Zeng Y, Xie H, Gu F, Chen J, Zhang G, Jiang X, et al. Artificial intelligence agent for delirium screening among patients in oncology and cardiac intensive care units: a proof-of-concept study. European Journal of Cardiovascular Nursing. 2026; 25: 810–817. https://doi.org/10.1093/eurjcn/zvag029

[49] Koizia LJ, Allott VES, Harris BHL. Detecting delirium with agents: implications of AI enabled screening for clinical care. European Journal of Cardiovascular Nursing. 2026; 25: 818–819. https://doi.org/10.1093/eurjcn/zvag033

[50] Lu J, Huang J, Guo Y, Wu Q, Jiang Z, Yang T, et al. Comparative analysis of six large language models in perioperative decision support for geriatric patients with multimorbidity: a three-dimensional evaluation framework. BMC Anesthesiology. 2026; 26: 88. https://doi.org/10.1186/s12871-025-03605-x

[51] Yu E, Chu X, Zhang W, Meng X, Yang Y, Ji X, et al. Large Language Models in Medicine: Applications, Challenges, and Future Directions. International Journal of Medical Sciences. 2025; 22: 2792–2801. https://doi.org/10.7150/ijms.111780

[52] Zeng Z, Hsu CL, Sit CHP, Wong SHS, Yang Y. The Role of Physical Activity and Physical Function in Predicting Physical Frailty Transitions in Chinese Older Adults: Longitudinal Observational Study From CHARLS. JMIR Aging. 2025; 8: e75887. https://doi.org/10.2196/75887

[53] Greco M, Angelucci A, Avidano G, Marelli G, Canali S, Aceto R, et al. Wearable Health Technology for Preoperative Risk Assessment in Elderly Patients: The WELCOME Study. Diagnostics (Basel, Switzerland). 2023; 13: 630. https://doi.org/10.3390/diagnostics13040630

[54] Song A, Xu K, Wang KY, Kong L, Kim DH. Detecting Pre-Frailty or Frailty Using Activity Monitoring from Wearable Sensor. Innovation in Aging. 2025; 9. https://doi.org/10.1093/geroni/igaf122.4315

[55] Osuka Y, Chan LLY, Brodie MA, Okubo Y, Lord SR. A Wrist-Worn Wearable Device Can Identify Frailty in Middle-Aged and Older Adults: The UK Biobank Study. Journal of the American Medical Directors Association. 2024; 25: 105196. https://doi.org/10.1016/j.jamda.2024.105196

[56] Tsukakoshi D, Yamamoto S, Nojima I, Sato M, Furuhashi K, Takeda S, et al. Association between postoperative delirium and heart rate variability in the intensive care unit and readmissions and mortality in elderly patients with cardiovascular surgery. Heart and Vessels. 2023; 38: 438–447. https://doi.org/10.1007/s00380-022-02173-1

[57] Halimeh M, Yang Y, Sheehan T, Vieluf S, Jackson M, Loddenkemper T, et al. Wearable device assessments of antiseizure medication effects on diurnal patterns of electrodermal activity, heart rate, and heart rate variability. Epilepsy & Behavior : E&B. 2022; 129: 108635. https://doi.org/10.1016/j.yebeh.2022.108635

[58] He C, Yu X, Zhang Y, Li Y, Jiang N. Wearable EEG devices in the detection of mild cognitive impairment: a systematic review. NPJ Digital Medicine. 2026; 9: 265. https://doi.org/10.1038/s41746-026-02342-w

[59] Xu A, Arnaout B, Scott DA, McGuigan S. Intraoperative electroencephalogram-derived measures and their association with postoperative delirium: a systematic review and meta-analysis. British Journal of Anaesthesia. 2025; 135: 1684–1703. https://doi.org/10.1016/j.bja.2025.08.038

[60] Butris N, Tang E, Pivetta B, He D, Saripella A, Yan E, et al. The prevalence and risk factors of sleep disturbances in surgical patients: A systematic review and meta-analysis. Sleep Medicine Reviews. 2023; 69: 101786. https://doi.org/10.1016/j.smrv.2023.101786

[61] Leung JM, Tang C, Do Q, Sands LP, Tran D, Lee KA, et al. Sleep Loss the night before surgery and incidence of postoperative delirium in adults 65-95 years of age. Sleep Medicine. 2023; 105: 61–67. https://doi.org/10.1016/j.sleep.2023.03.015

[62] Pandrangi VC, Jorizzo M, Shah S, Bruening J, Wax MK, Clayburgh D, et al. Monitoring postoperative ambulation and sleep after head and neck surgery: Feasibility and utility study using wearable devices. Head & Neck. 2022; 44: 2744–2752. https://doi.org/10.1002/hed.27182

[63] Angelucci A, Greco M, Cecconi M, Aliverti A. Wearable devices for patient monitoring in the intensive care unit. Intensive Care Medicine Experimental. 2025; 13: 26. https://doi.org/10.1186/s40635-025-00738-8

[64] Bruzzone M, Chapin B, Arias F, Walker J, Price CC. A systematic review on the perioperative use of EEG as a predictor/diagnostic tool for postoperative delirium. Alzheimer's & Dementia. 2023; 19: e075376. https://doi.org/10.1002/alz.075376

[65] Bruzzone MJ, Chapin B, Walker J, Santana M, Wang Y, Amini S, et al. Electroencephalographic Measures of Delirium in the Perioperative Setting: A Systematic Review. Anesthesia and Analgesia. 2025; 140: 1127–1139. https://doi.org/10.1213/ANE.0000000000007079

[66] Ning MH, Rodionov A, Ross JM, Ozdemir RA, Burch M, Lian SJ, et al. Prediction of postoperative delirium in older adults from preoperative cognition and occipital alpha power from resting-state electroencephalogram. Age and Ageing. 2025; 54: afaf330. https://doi.org/10.1093/ageing/afaf330

[67] Hata M, Miyazaki Y, Nagata C, Masuda H, Wada T, Takahashi S, et al. Predicting postoperative delirium after cardiovascular surgeries from preoperative portable electroencephalography oscillations. Frontiers in Psychiatry. 2023; 14: 1287607. https://doi.org/10.3389/fpsyt.2023.1287607

[68] Aldecoa C, Bettelli G, Bilotta F, Sanders RD, Aceto P, Audisio R, et al. Update of the European Society of Anaesthesiology and Intensive Care Medicine evidence-based and consensus-based guideline on postoperative delirium in adult patients. European Journal of Anaesthesiology. 2024; 41: 81–108. https://doi.org/10.1097/EJA.0000000000001876

[69] Likhvantsev VV, Berikashvili LB, Smirnova AV, Polyakov PA, Yadgarov MY, Gracheva ND, et al. Intraoperative electroencephalogram patterns as predictors of postoperative delirium in older patients: a systematic review and meta-analysis. Frontiers in Aging Neuroscience. 2024; 16: 1386669. https://doi.org/10.3389/fnagi.2024.1386669

[70] Röhr V, Blankertz B, Radtke FM, Spies C, Koch S. Machine-learning model predicting postoperative delirium in older patients using intraoperative frontal electroencephalographic signatures. Frontiers in Aging Neuroscience. 2022; 14: 911088. https://doi.org/10.3389/fnagi.2022.911088