BJA

British Journal of Anaesthesia, xxx (xxx): xxx (xxxx)

doi: 10.1016/j.bja.2026.01.025
Advance Access Publication Date: xxx
Clinical Investigation

CLINICAL INVESTIGATION

Prediction and risk evaluation of delirium after surgery in older patients: development and internal validation of an algorithm from the prospective BioCog cohort study

Florian Lammers-Lietz[1,2,*], Levent Akyuez[3,4], Diana Boraschi[5], Friedrich Borchers[1], Jeroen de Bresser[6], Sreyoshi Chatterjee[7,8], Marta M. Correia[9], Nikola M. de Lange[7], Thomas Bernd Dschietzig[10], Soumyabrata Ghosh[7], Insa Feinkohl[11,12,‡], Izabela Ferreira da Silva[7], Marinus Fislage[1], Anna Fournier[7,13], Jürgen Gallinat[14], Daniel Hadzidiakos[1], Sven Hädel[2], Fatima Halzl-Yürek[1], Stefanie Heilmann-Heimbach[15], Maria Heinrich[1,16], Jeroen Hendrikse[17], Per Hoffmann[15,18,19], Jürgen Janke[11,20], Ilse M. J. Kant[21], Angelie Kraft[22,23], Roland Krause[7], Jochen Kruppa-Scheetz[24,25], Simone Kühn[1,14], Gunnar Lachmann[1,16], Markus Laubach[1,23,26], Christoph Lippert[27], David K. Menon[28,29], Rudolf Mörgeli[1], Anika Müller[1], Henk-Jan Mutsaerts[30,31], Markus Nöthen[15], Peter Nürnberg[32,33,34], Kwaku Ofosu[1], Malte Pietzsch[35], Sophie K. Piper[24,36], Tobias Pischon[11,20,37], Jacobus Preller[38,39], Konstanze Scheurer[1], Reinhard Schneider[7], Kathrin Scholtz[1], Peter H. Schreier[23,33], Arjen J. C. Slooter[21,§], Emmanuel A. Stamatakis[29,40], Clarissa von Haefen[1], Simone J. T. van Montfort[21], Edwin van Dellen[21,41], Hans-Dieter Volk[4,42], Simon Weber[35], Janine Wiebach[24,43], Anton Wiehe[2,22,23], Jeanne M. Winterer[2,23,44,45], Alissa Wolf[1], Norman Zacharias[1,23,46], Claudia Spies[1,†], Georg Winterer[1,2,23,†] on behalf of the BioCog consortium

[1]Department of Anesthesiology and Intensive Care Medicine—Universitätsmedizin Berlin, Berlin, Germany, [2]Pharmaimage Biomarker Solutions GmbH, Berlin, Germany, [3]Berlin Institute of Health at Charité—Universitätsmedizin Berlin, BIH Center for Regenerative therapies (BCRT), and Charité—Universitätsmedizin Berlin, Institute of Medical Immunology, Berlin, Germany, [4]CheckImmune GmbH, Berlin, Germany, [5]Institute of Protein Biochemistry, Consiglio Nazionale delle Ricerche (CNR) di Pisa, Pisa, Italy, [6]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands, [7]Bioinformatics Core, Luxembourg Center for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg, [8]Deutsches Zentrum für Luft und Raumfahrt (DLR), Cologne, Germany, [9]MRC Cognition and Brain Sciences Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK, [10]Immundiagnostik AG, Bensheim, Germany, [11]Molecular Epidemiology Research Group, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany, [12]Faculty of Health at Department of Medicine, Witten/Herdecke University, North Rhine-Westphalia, Germany, [13]Swiss Data Science Center (SDSC), ETH Zurich, Zurich, Switzerland, [14]Department of Psychiatry, University Medical-Center Hamburg-Eppendorf, Hamburg, Germany, [15]Institute of Human Genetics, University of Bonn, Bonn, Germany, [16]Berlin Institute of Health at Charité—Universitätsmedizin Berlin, BIH Academy, (Digital) Clinician Scientist Program, Berlin, Germany, [17]Department of Radiology and Brain Center Rudolf Magnus, University Medical Center Utrecht (UMC), Utrecht, The Netherlands, [18]Division of Medical Genetics, University Hospital, Basel, Switzerland, [19]Human Genetics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland, [20]Biobank Technology Platform, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany, [21]Department of Intensive Care Medicine and Brain Center, University Medical Center Utrecht (UMC), Utrecht University, Utrecht, The Netherlands, [22]AdaLab UG, Hamburg, Germany, [23]Pharmaimage

Received: 28 April 2025; **Accepted:** 13 January 2026

Biomarker Solutions Inc., Cambridge, MA, USA, [24]Institute of Biometry and Clinical Epidemiology—Universitätsmedizin Berlin, Berlin, Germany, [25]Hochschule Osnabrück, University of Applied Sciences, Osnabrück, Germany, [26]Department of Orthopaedics and Trauma Surgery, Musculoskeletal University Center Munich (MUM), LMU University Hospital, LMU Munich, Munich, Germany, [27]Hasso-Plattner Institute, University of Potsdam, Potsdam, Germany, [28]Neurosciences/Trauma Critical Care Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, [29]Division of Anaesthesia, Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge, UK, [30]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, The Netherlands, [31]Amsterdam Neuroscience, Amsterdam, The Netherlands, [32]Cologne Center for Genomics, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany, [33]Institute for Genetics of the University of Cologne, Cologne, Germany, [34]Atlas Biolabs GmbH, Berlin, Germany, [35]Cellogic GmbH (Cellogic), Berlin, Germany, [36]Institute of Medical Informatics—Universitätsmedizin Berlin, Berlin, Germany, [37]Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Core Facility Biobank, Berlin, Germany, [38]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge, UK, [39]John Farman Intensive Care Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, [40]Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge, UK, [41]Department of Psychiatry and UMC Utrecht Brain Center, University Medical Center Utrecht (UMC), Utrecht, The Netherlands, [42]Institute of Medical Immunology, Charité—Universitätsmedizin Berlin—Universitätsmedizin Berlin, Berlin, Germany, [43]BIH Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Core Unit Metabolomics, Berlin, Germany, [44]Department of Psychiatry—Universitätsmedizin Berlin, Berlin, Germany, [45]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany and [46]Department of Otorhinolaryngology, Head and Neck Surgery, Charité—Universitätsmedizin Berlin, Berlin, Germany

*Corresponding author. E-mail: florian.lammers@charite.de
[†]Equally contributing senior authors.
[‡]Current address: Epidemiology Research Group, Faculty of Health Sciences Brandenburg, University of Potsdam, Potsdam, Germany.
[§]Current address: Department of Psychiatry, University Medical Center Groningen and Department of Psychiatry, and Research School of Behavioural and Cognitive Neurosciences (BCN), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

Abstract

Background: Postoperative delirium (POD) affects ~20% of older surgical patients. It is associated with poor clinical outcome and increased mortality. We aimed to identify the major POD risk factors and to develop and validate a multivariate algorithm for individual POD risk prediction and risk evaluation in the very early postoperative period.
Methods: BioCog is a prospective cohort study conducted in the anaesthesiology departments of two tertiary care centres in Germany and The Netherlands. Patients aged ≥65 yr with no preoperative dementia (Mini-Mental Status Examination ≥24) undergoing surgery with an expected duration of at least 60 min were enrolled and screened for POD according to DSM 5 until the seventh postoperative day. Clinical, neuropsychological, neuroimaging data, and blood were measured before and after surgery. We evaluated several models by sequentially adding blocks of variables. Gradient-boosted trees (GBT) with nested cross-validation were used for POD prediction. Model accuracy (area under the receiver-operating curve, AUC) and calibration were assessed (Brier score).
Results: Out of 929 patients, 184 (20%) experienced POD. A GBT algorithm using both preoperative data, characteristics of the intervention, and postoperative changes in laboratory parameters achieved the highest AUC (0.83, [0.79—0.86]) with a Brier score of 0.12 (0.12—0.13).
Conclusions: Models combining preoperative with precipitating factors during surgery predict POD with high accuracy. This suggests that the resulting algorithms eventually may become useful to support clinical decision-making.
Clinical trial registration: NCT02265263.

Keywords: cohort study; neuroimaging; postoperative complications; postoperative delirium; risk factors; transcriptome

Editor's key points

- Postoperative delirium (POD) is a frequent and harmful complication of surgery.
- Using data from a prospective cohort study, the BioCog consortium applied machine learning models to predict POD.
- Gradient-boosted tree models integrating clinical, molecular, and neuroimaging data from 929 patients aged ≥65 yr undergoing elective surgery showed that preoperative data, details of the intervention, and postoperative changes in laboratory parameters can achieve good predictive performance (AUC ≥0.80).
- The data underscore the importance of procedure-related factors in development of POD.

Delirium is an acute disturbance in attention, awareness, cognition, psychomotor behaviour, and emotional state because of another medical condition. The incidence of postoperative delirium (POD) ranges 5–50%,[1] but it is most frequent in older patients.[2,3] POD incidence is associated with poor cognitive outcomes, hospitalisation, re-institutionalisation, increased costs, and mortality.[3,4]

Various previous studies have tried to build machine learning–based prediction tools for POD, usually based on retrospective analyses.[5–9] Only two prospective studies achieved area under the receiver-operating curve (AUC) values of 0.71–0.74.[10,11] The prospective Biomarker Development for Postoperative Cognitive Impairment in the Elderly (BioCog) study was conducted with the main goal to improve POD estimation. We were taking a systems medicine approach with focus on inflammatory alterations and the immune system, the cholinergic system and metabolic changes, and indicators for early dementia based on an in-depth systematic review.[1] Investigations included a wide range of perioperative clinical and neuropsychological parameters, neuroimaging, laboratory investigations, and gene expression. Furthermore, the incorporation of precipitating factors may have additional value to predisposing factors.

The primary aim of this study was to develop and internally validate a POD risk index based on multimodal non-routine data intended for use by healthcare professionals to advise patients during medical decision-making and allocating healthcare resources both during surgery planning in the preoperative phase and in the immediate postoperative phase up to the first postoperative day.

Methods

Study design

BioCog (clinicaltrials.gov: NCT02265263) is a prospective observational cohort study with the aim of identifying POD risk factors. The model was developed and internally validated in this cohort. All procedures were approved by the local ethics committees in Berlin, Germany (EA2/092/14) and Utrecht, The Netherlands (14-469), and conducted in line with the declaration of Helsinki. All participants gave written informed consent before inclusion.

Participants

Male and female patients were enrolled in two tertiary care centres at the Charité–Universitätsmedizin Berlin, Berlin Germany, and the University Medical Center, Utrecht, The Netherlands. Consenting patients aged ≥65 yr presenting for elective surgery with an expected duration of >60 min were included. Patients meeting one of the following criteria were excluded: (1) positive screening for pre-existing major neurocognitive disorder defined as a Mini-Mental Status Examination (MMSE) score ≤23 points; (2) any condition interfering with neurocognitive assessment (severe sensory impairment, neuropsychiatric illness including alcohol and drug dependence, intracranial surgery); (3) unavailability for follow-up assessment; (4) accommodation in an institution owing to official or judicial order; and (5) inability to give informed consent.

Study procedures

The preoperative data were collected at least 1 day before surgery, including medical history and clinical assessments, neuropsychological testing, blood collection, and neuroimaging. Postoperative study visits took place twice daily until the seventh postoperative day.

Outcome

POD during the first 7 days after surgery was the primary endpoint. Independently of the routine hospital procedures, POD screening was started in the recovery room and repeated twice per day at 8:00 and 19:00 (±1 h) up to 7 days after surgery, by or under the supervision of a study physician. POD was defined according to Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) criteria and assessed by prospective screening with three validated tools which were recorded at each visit in accordance with current guidelines,[2,3] to mitigate the known tendency of physicians to underdiagnose POD. Prospective screening was supplemented by chart review for delirium to account for symptom fluctuations in delirious patients between study visits, especially during night.[12] Patients were considered delirious if at least one of the following criteria was positive: (1) ≥2 points on the Nursing Delirium Screening Scale (Nu-DESC); (2) positive Confusion Assessment Method (CAM) score on a general ward; (3) positive CAM for the Intensive Care Unit (CAM-ICU) score on an ICU; and (4) chart review (e.g. ward nurses and physician notes) showing descriptions of delirium.

Clinical assessments

Before surgery, the study team recorded sociodemographic data and information on medication according to Carnahan's Anticholinergic Drug Scale, health-related quality of life (EQ5D), Mini-Nutritional Assessment (MNA) and BMI, tobacco and hazardous alcohol consumption (Alcohol Use Disorders Identification Test [AUDIT]). A functional and physical assessment battery including frailty and walking speed was conducted. Precipitating factors recorded were duration of surgery and anaesthesia, type of anaesthetic procedure (regional anaesthesia, general anaesthesia, or both), type of surgery (intracranial, intrathoracic/-abdominal/-pelvic surgery, or peripheral), postoperative pain, prescription of

anticholinergic medication daily until the seventh postoperative day, length of hospital and ICU stay, and complications and postoperative mortality until the 90th postoperative day (Supplementary Chapters 1.1–1.6).

Neuropsychological data

The preoperative cognitive assessment consisted of a comprehensive screen-based neuropsychological test battery (CANTAB; Cambridge Cognition Ltd., Cambridge, UK) and additional tests (Trail-Making-Test Parts A and B, and Grooved Pegboard Test). MMSE score at the screening visit, CANTAB test scores, and overall preoperative cognitive impairment (PreCI) were analysed as risk factors.

PreCI is a dichotomous variable defined through comparison of cognitive test performance with a control group. We used multiple cognitive test parameters with moderate-to-good retest–reliability in the control group and calculated z-scores of the baseline measurement in each test parameter assessed in the control group.[13] The same z-transformation was then applied to the surgical cohort. Z-scores <−1.96 in at least two cognitive test parameters or an averaged z-score <−1.96 was used to define PreCI (Supplementary Table 2).

Laboratory parameters

Preoperative serum and plasma samples were collected in supine position immediately before induction of anaesthesia after 8 h of fasting and on the morning of the first postoperative day. Blood sampling was performed by trained clinic staff according to a standard operating procedure adapted from the German National Cohort Study.[14] Samples were immediately sent to laboratories adjacent to the respective hospital site for analysis, or frozen at −80°C and shipped to a central biobank for sample processing and storage (Supplementary Chapter 1.8). Whenever necessary, values were adjusted for laboratory.

Transcriptomics

Samples for transcriptomic analysis were collected in PAXgene tubes (Qiagen, Hilden, Germany) at the same time points as other blood samples. Analyses were performed with Affymetrix Clariom S human microarray for RNA and Affymetrix® Flash Tag™ Biotin HSR (miRNA 4.1 Array Plates) for microRNA analyses (Thermo Fischer, Santa Clara, CA, USA) in a GeneTitan™ Multi-Channel Instrument by Atlas Biolabs GmbH (Berlin, Germany).

Neuroimaging

The MRI protocol included whole brain T1- and T2-weighted high-resolution hippocampus imaging and diffusion tensor imaging (DTI). In addition, functional MRI and arterial spin labelling were conducted, but have not been considered for prediction owing to low between-scanner agreement (interclass correlation coefficient of 0.36–0.54 for functional connectivity in default mode, salience executive, and dorsal attention networks, and 0.17–0.39 for quantified cerebral blood flow). We calculated global and regional brain volumes including hippocampal subregions, cortical thickness, and curvature from T1-weighted imaging, mean diffusivity, kurtosis, and fractional anisotropy from DTI (Supplementary Chapter 1.10).

Statistical analysis

Estimation of sample size

The rule of thumb of Harrell was used to plan an appropriate number of POD events for a stable prediction model (i.e. ≥10 events per independent variable in logistic regression), which was considered adequate for machine learning.[15] Requiring 260 patients with POD for analysis of up to 26 independent predictor variables and expecting a 25% incidence of POD, the number of required patients was 1040. Assuming a dropout rate of 15%, a total number of 1200 patients was planned. The initial analysis plan stipulated a training/test split approach for internal validation owing to its computational efficiency. As the study finally achieved a lower cohort size (n=929), nested k-fold cross-validation was used instead which works more efficiently on small samples.

Analysis of single parameters

For descriptive purposes, associations of preoperative/perioperative parameters with POD were analysed using simple logistic regression. We report odds ratios (ORs) with 95% confidence intervals (CIs) for the depending variable POD (reference category: no POD). To improve the interpretability of single parameter analyses, standardising transformations were applied to the raw variables (Supplementary Chapter 1.11.1) so that the OR refers to a change by one standard deviation on the independent variable, or dichotomised according to clinically relevant cut-off values for presentation of interpretable ORs. Standardisation and use of cut-off values are reported in the results section, if applicable. Analyses were conducted in R v3.5 (R Foundation for Statistical Computing, Vienna, Austria) and SPSS software (IBM, Armonk, NY, USA). No adjustments for multiple testing were made; therefore, results should be considered exploratory, and we abstain from reporting P-values.

Machine learning

We applied machine learning (gradient-boosted trees [GBT]) to explore how the interplay of a larger set of predictors would benefit the prediction of POD risk in a bottom-up, data-driven fashion. Data available before surgery or on the first postoperative day were eligible for inclusion in machine learning, as these data were deemed useful for preoperative POD risk prediction and postoperative evaluation. Please refer also to Supplementary Chapter 1.11.2 for additional implementation details not given in the following text.

Variables were assembled into blocks: preoperative data from the clinical assessment ('Clinical'), characteristics of the surgical intervention ('Precipitants', 'Pain'), preoperative neuroimaging data ('Imaging'), preoperative values and perioperative difference in blood parameters ('Blood' and 'Blood periop.'), preoperative RNA and µRNA abundance ('RNA' and 'µRNA'), and perioperative difference in transcript abundance ('RNA periop.'). Perioperative differences here refer to the difference between the value measured on the morning of the first postoperative day and the preoperative values. Different GBT models were built on combinations of various variable blocks. Combinations were selected sequentially, starting with simple models (i.e. using only variables from one block) and then adding further blocks based on the AUC, assumptions on feasibility, and relevance for clinical routine.

Table 1 Sample description (N=929). *End of anaesthesia was assessed differentially in both study centres. †Relative frequencies are calculated after correction for missing values. ‡Length of survival until the 90th postoperative day is only available for the study centre in Berlin. In Utrecht, three patients were deceased before follow-up assessment after 3 months and 27 did not attend follow-up for other reasons. ¶Complications were recorded until discharge in Berlin, and until the seventh postoperative day in Utrecht. §Intracranial surgery not affecting brain parenchyma (e.g. meningioma). ASAT, aspartate aminotransferase; AUDIT, Alcohol Use Disorders Identification Test; CRP, C-reactive protein; freq., frequency; GDS, geriatric depression scale; GGT, γ-glutamyltransferase; IQR, interquartile range; ISCED, International Standard Classification for Education; LDL, low-density lipoprotein; max, maximum; min, minimum; MMSE, Mini-Mental Status Examination; MNA, Mini-Nutritional Assessment; NBM, nucleus basalis of Meynert; NT-proBNP, N-terminal pro-brain natriuretic peptide; PreCI, preoperative cognitive impairment; preop., preoperative; POD, postoperative delirium; postop., postoperative, referring to measurements on the first postoperative day for blood-based variables.

		All		POD		No POD	
		Median (IQR)	Min–max	Median (IQR)	Min–max	Median (IQR)	Min–max
Age (yr)		72 (69–76)	65–91	74 (71–73)	65–91	71 (68–75)	65–90
BMI (kg m^{-2})		26.6 (24.0–29.4)	14.7–46.8	26.6 (23.7–29.7)	17.6–44.3	26.6 (24.2–29.4)	14.7–46.8
MMSE score (points)		29 (28–30)	24–30	28 (27–30)	24–30	29 (28–30)	24–30
GDS		1 (0–3)	0–13	2 (1–3)	0–13	1 (0–3)	0–11
EQ5D		0.88 (0.76–1.00)	−0.14 to 1.00	0.86 (0.68–1.00)	0.17–1.00	0.89 (0.78–1.00)	−0.14 to 1.00
Charlson's comorbidity index (p)		1 (0–2)	0–10	2 (0–3)	0–7	1 (0–2)	0–10
Preop. haemoglobin (g dl^{-1})		13.1 (11.9–14.3)	5.4–17.9	12.5 (11.0–13.8)	7.0–16.7	13.3 (12.1–14.3)	5.4–17.9
Postop. haemoglobin (g dl^{-1})		11.7 (10.0–13.0)	5.8–16.0	10.3 (8.9–11.9)	5.8–15.6	11.9 (10.4–13.1)	5.8–16.0
Preop. CRP (mg L^{-1})		3.4 (1.4–8.3)	0.1–232.0	5.7 (2.3–11.2)	0.3–105.0	3.0 (1.2–7.3)	0.1–232.0
Postop. CRP (mg L^{-1})		34.4 (5.9–56.8)	0.1–253	51.6 (41.2–91.7)	0.8–227	26.6 (5.0–53.5)	0.1–253
Preop. leucocytes (nl^{-1})		6.2 (5.0–7.5)	1.6–24.6	6.3 (4.9–7.8)	2.7–19.4	6.2 (5.0–7.4)	1.6–24.6
Postop. leucocytes (nl^{-1})		9.5 (7.8–11.8)	2.6–31.0	10.4 (8.5–12.8)	4.6–31.0	9.4 (7.5–11.5)	2.6–24.6
Preop. IL6 (pg ml^{-1})		2.0 (0.0–5.0)	0.0–423.7	2.7 (0.9–6.8)	0.0–369.7	1.7 (0.0–4.5)	0.0–423.7
Postop. IL6 (pg ml^{-1})		40.0 (13.5–123.1)	0.0–468.0	111.8 (35.9–259.1)	0.0–461.2	32.0 (11.7–91.3)	0.0–468.0
Preop. albumin (g L^{-1})		40.7 (37.8–43.2)	15.5–51.7	38.6 (35.3–42.3)	22.1–51.7	41.1 (38.4–43.3)	15.5–51.6
Preop. creatinine (μM)		76.0 (64.5–90.2)	32.7–529.5	75.1 (65.2–89.5)	35.4–529.6	76.0 (64.5–90.6)	32.7–414.6
Preop. NT-proBNP (pM)		6.1 (2.9–21.4)	2.9–617.2	7.1 (2.9–23.4)	2.9–186.7	6.0 (2.9–20.2)	2.9–617.2
Preop. LDL cholesterol (mM)		3.0 (2.3–3.7)	0.1–7.7	2.9 (2.2–3.5)	0.1–6.4	3.0 (2.4–3.6)	0.5–7.7
Brain volume (cm^3)		1000 (927–1086)	721–1424	961 (910–1046)	721–1208	1007 (932–1089)	721–1424
NBM volume (mm^3)		1752 (1623–1878)	1287–2321	1716 (1612–1833)	1287–2138	1759 (1626–1884)	1287–2321
Hippocampus volume (mm^3)		3579 (3322–3852)	2114–4448	3426 (3190–3684)	2681–4348	3595 (3368–3868)	2114–4448
Duration of anaesthesia (Utrecht, min)*		265 (213–390)	10–1669	314 (238–541)	12–1663	260 (211–368)	10–1669
Duration of anaesthesia (Berlin, min)*		167 (106–279)	25–753	334 (202–470)	55–753	150 (98–223)	25–738
Duration of surgery (Berlin, min)†		102 (55–191)	3–594	220 (104–360)	18–594	89 (47–148)	3–572
Duration of hospital stay (days)		7 (4–11)	1–131	11 (7–23)	1–131	5 (3–8)	1–69
Duration of ICU stay (days)		0 (0–0)	0–55	1 (0–3)	0–55	0 (0–0)	0–22
		Absolute *n*†	**Relative freq.† (%)**	**Absolute *n*†**	**Relative freq.† (%)**	**Absolute *n*†**	**Relative freq.† (%)**
PreCI		122/924	13	43/183	23	79/741	11
Mortality at 3 months (Berlin)‡		29/683	4	19/141	13	10/542	2
Complications¶	Death	18/921	2	5/184	3	13/737	2
	Nonfatal	457/921	50	161/184	88	296/737	40
Site of surgery	Intracranial§	10/911	1	2/179	1	8/732	1
	Intrathoracic, -abdominal, -pelvic	397/91	44	116/179	65	281/732	38
	Peripheral	505/911	55	61/179	34	443/732	61
Type of anaesthesia	General	687/912	75	122/170	72	565/732	77
	Regional	57/912	6	4/170	2	53/732	7
	Combined	168/912	18	54/170	32	114/732	16
Benzodiazepine long-term intake		173/922	19	60/184	33	113/744	15

Continued

Table 1 Continued

		All		POD		No POD	
		Median (IQR)	Min–max	Median (IQR)	Min–max	Median (IQR)	Min–max
Benzodiazepine premedication		108/874	12	25/166	15	83/708	12
Postop. pain	Screening period	335/904	37	93/179	52	242/725	33
	Day of surgery	248/712	35	49/126	39	199/586	34
ASA physical status	1	36/929	4	3/184	2	33/745	4
	2	557/929	60	81/184	44	476/745	64
	3	335/929	36	100/184	54	235/745	32
	4	1/929	<1	0/184	0	1/745	<1
Women		394/929	42	85/394	22	99/535	19
MNA	Normal	662/911	73	111/182	61	551/729	76
	At risk	200/911	22	55/182	30	145/729	20
	Malnourishment	49/911	5	16/182	9	33/729	5
Frailty (Fried)	Robust	354/631	56	56/142	39	298/489	61
	Prefrail	175/631	28	47/142	33	128/489	26
	Frail	102/631	16	39/142	27	63/489	13
Smoker		90/903	10	18/175	10	72/728	10
Hazardous alcohol consumption (AUDIT)		62/862	7	14/168	8	46/694	7
ISCED 1997	Level 1+2	150/839	18	29/165	18	121/674	18
	Level 3+4	343/839	41	64/165	39	279/674	41
	Level 5+6	346/839	41	72/165	44	274/674	41

The GBT algorithm takes a set of decision trees as weak classifiers and combines them to form a strong classifier. It does so by incrementally adding decision trees during training to steadily improve its previous performance. The sampling of input cases is focused on those cases that were hard to classify before training and individual tree predictions are weighted.

The GBT models handle missing values natively during training. At each decision node, the algorithm learns an optimal default direction for samples with missing values by evaluating which branch yields the highest information gain. This allows the model to utilise all available data without requiring explicit imputation. We provide model performance parameters for both the full cohort including missing values and a complete-case analysis.

During inference, the output is computed through sequential application of each tree. GBT provides a continuous output parameter bounded between 0 and 1, allowing the choice of a clinically relevant cut-off which can be flexibly adapted to address various clinical questions and is inherently able to handle missing data. Algorithms were programmed in Python (Python Software Foundation, Wilmington, DE, USA), using the GBT implementation of the XGBoost library (RRID:SCR_021361) with a logistic loss function (binary cross-entropy). GBT models are inherently dependent on several key hyperparameters. To optimise these hyperparameters, we defined specific value ranges based on plausible expectations, establishing a search space anticipated to contain the optimal settings for our objectives (maximum depth: integer range 3–10; learning rate: floating range 0.005–0.1 on a logarithmic scale; number of estimators: integer range 5–100; subsample: floating range from 0.8–1.0; column sample by tree: floating range 0.6–1.0; Γ: floating range 0.0–5.0). Hyperparameter optimisation was automated using a Bayesian search strategy (Optuna [Preferred Networks Inc., Tokyo, Japan]) to efficiently find the optimal parameter set within the defined search space. Models were validated using nested cross-validation. This approach allows model hyperparameter optimisation and model selection while avoiding model overfitting. Although each of the training datasets is provided to a hyperparameter optimised procedure, the evaluation of hyperparameters is performed using another cross-validation procedure that splits up each of the provided train dataset into another set of k-folds.

The AUC and the Brier score with 95% CI are provided. Model performance was primarily assessed by the AUC-ROC. To evaluate the reliability of the probabilistic predictions, we also calculated the Brier score and generated calibration and binned residual plots, which assess the agreement between predicted probabilities and actual outcomes. The Brier score measures the difference between predicted probabilities and actual outcomes, ranging between 0, for perfect prediction, and 1.

Sex-specific analyses have been conducted for the best-performing model.

Results

We recruited 933 patients between November 2014 and April 2017. Table 1 characterises the sample. The patient flow chart is given in Figure 1. Additional details on excluded patients are given in Supplementary Chapter 2.1. POD assessments were available for 929 patients. Of these, 184 (20%) patients developed POD (141/686 [21%] in Berlin, Germany and 43/243 [18%] in Utrecht, The Netherlands).

Out of 184 patients, 83 (45%) patients with POD were identified in the bedside screening only, 13 (7%) were diagnosed from chart review only, and 88 (48%) were proved in both chart review and bedside screening (Supplementary Fig. 4, Supplementary Table 3).

Furthermore, 374 patients in the cohort stayed for a period of 8 days or longer in hospital, and 361 patients were screened for delirium on the seventh postoperative day.

Postoperative delirium risk factors

Figure 2 displays unadjusted OR with 95% CIs for preoperative parameters with CIs excluding unity. Sample and effect sizes for all parameters are given in the Supplementary material.

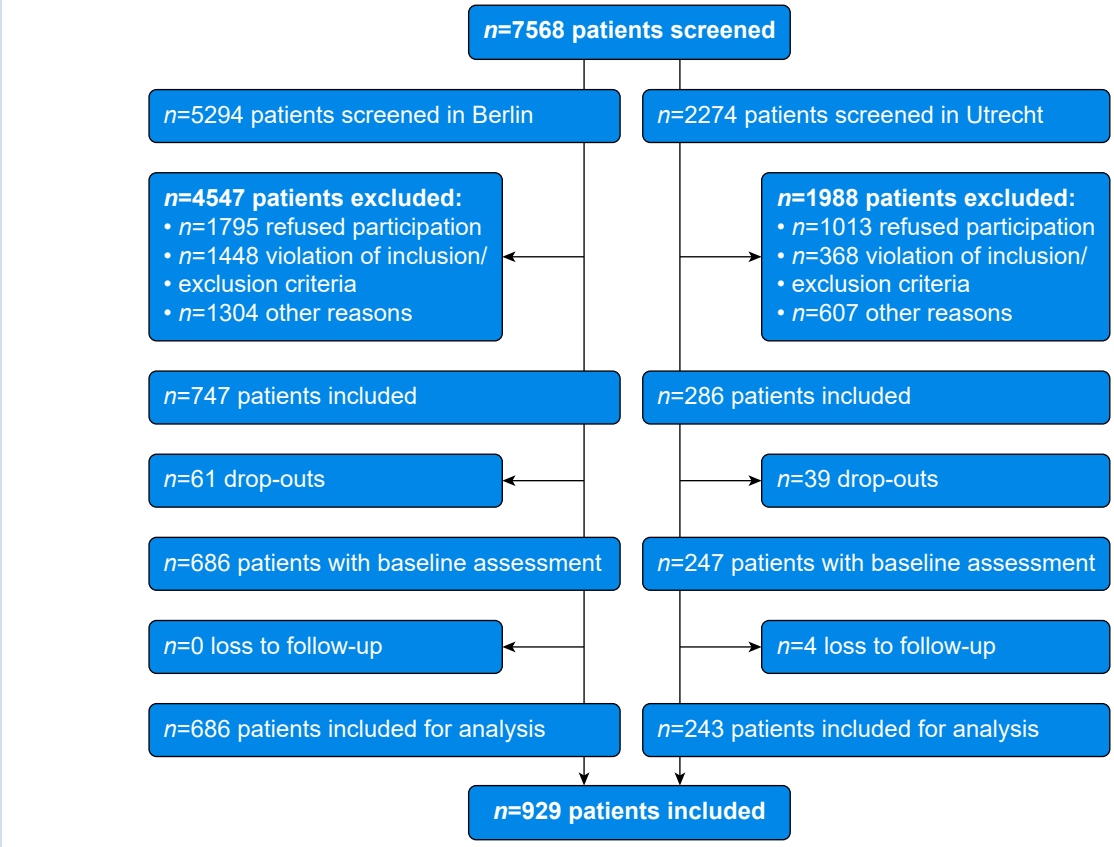

Fig 1. Patient flow chart.

Age was directly associated with POD. Among age-related conditions, frailty had the strongest association with POD, and so were slow walking speed, malnutrition, any functional impairment according to Instrumental Activities of Daily Living or Barthel index and depressive symptoms. An MMSE score <27points had a higher OR for POD than PreCI (Supplementary Fig. 5 and Supplementary Table 8). Preoperative higher concentrations of cholesterol and associated lipoproteins (HDL and LDL) were protective against POD. A postoperative decrease in triglycerides, cholesterol, and LDL were associated with higher POD incidence.

Four inflammatory parameters were positively associated with POD: interleukin (IL)-6, whole blood IL8,[18] C-reactive protein (CRP), immature granulocyte fraction, and neutrophil count. An increase of inflammatory parameters on the first postoperative assessment was associated with higher likelihood of POD (CRP: standardised, adjusted OR 1.59 [1.14–2.21], IL6: standardised OR 1.76 [1.48–2.09], and IL8: standardised OR 1.96 [1.18–3.24]). Cellular immune response showed a more complex association with POD. Although a postoperative increase in leucocytes (standardised, adjusted OR 1.36 [1.12–1.64]) and neutrophils (standardised, adjusted OR 1.47 [1.2–1.81]) was associated with POD, an increase in lymphocytes lowered the odds for POD (standardised, adjusted OR 0.66 [0.54–0.81]).

Higher concentrations of tryptophan (standardised, laboratory-adjusted OR 0.74 [0.62–0.89]) and albumin also lowered the odds for POD. A higher plasma β-amyloid 42/40-ratio was found to be related to a lower POD likelihood, but this association seemed to be driven by increased POD risk in patients with higher concentrations of β-amyloid 40.

Both higher preoperative γ-glutamyltransferase concentrations and a postoperative decrease (standardised, adjusted OR 0.81 [0.68–0.98]) were associated with POD. A postoperative increase in transaminases was associated with POD. After surgery, decreasing concentrations of oxidative stress indicated by nitrotyrosine concentrations (standardised OR 0.72 [0.53–0.98]) and nitric oxide production indicated by homoarginine concentrations (standardised OR 0.48 [0.31–0.73]) were associated with increased POD risk.

Longer duration of anaesthesia and surgery and also blood loss (standardised, adjusted OR for perioperative changes in Hb: 0.76 [0.63–0.91], thrombocytes: 0.57 [0.46–0.69], and albumin: 0.66 [0.54–0.81]) were associated with POD. Compared with general anaesthesia, surgery performed in regional anaesthesia was associated with lower rates of POD. Surgery with opening of thorax, abdomen, or pelvis was associated with increased rates of POD compared with peripheral surgery (Supplementary Table 10). Pain or intake of any anticholinergic medication at least once during follow-up until the seventh postoperative day was associated with POD (Supplementary Figs 6 and 7, Supplementary Tables 11 and 12).

Various associations of structural MRI-derived parameters were observed (complete results: Supplementary Fig. 8, Supplementary Table 15). We would like to emphasise a

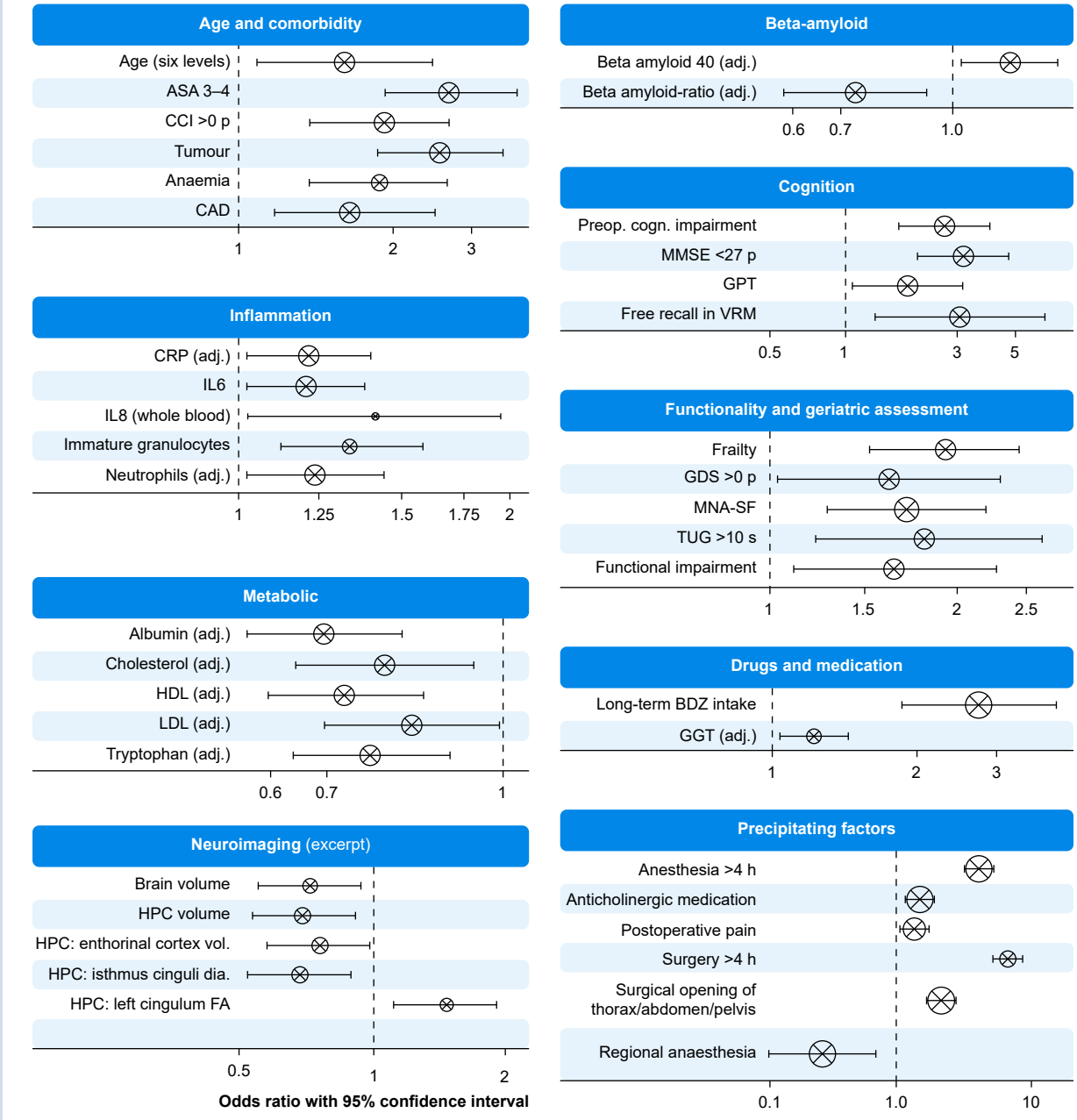

Fig 2. Summary of parameters that were significantly associated with postoperative delirium (POD). Odds ratios (OR) with 95% confidence interval (95% CI) are shown (only parameters are depicted with CI excluding unity). The diameter of the circle corresponds to the number of available datasets. If no information on cut-off values or categories is given, standardised ORs are given for continuously scaled variables. Age was split into six intervals of 5 yr for display purposes. For Mini-Nutritional Assessment–short form (MNA-SF) and frailty, previously described categories (normal nutritional status, risk of malnutrition, and malnourishment; robustness, pre-frailty, and frailty) have been aggregated from point scores as recommended.[16,17] The term tumour includes diagnoses of solid malignancies, leukaemia, and lymphoma. Functional impairment refers to presence of any functional impairment in either Barthel Index or index of activities of daily living. See also Supplementary material. Frailty refers to Fried's frailty phenotype. adj., adjusted for assessment in different study centres; BDZ, preoperative long-term prescription of benzodiazepines; CA, cornu ammonis; CAD, coronary artery disease; CCI, Charlson comorbidity index; CRP, C-reactive protein; dia., diameter (cortical thickness); FA, fractional anisotropy; GDS, Geriatric Depression Scale; GGT, γ-glutamyltransferase; GPT, Grooved Pegboard Test (completion time); HDL, high-density lipoprotein; HPC, hippocampus; IL, interleukin; LDL, low-density lipoprotein; MMSE, Mini-Mental Status Examination; p, points; preop. cogn. impairment, preoperative cognitive impairment; TUG, Timed up-and-go test; vol., volume; VRM, Verbal Recognition Memory.

protective association of POD with global brain volume and hippocampus volume.

Machine learning prediction of postoperative delirium

Figure 3 and Table 2 summarise model performances. Among the models using only preoperative data, the model using only clinical data performed best. Adding preoperative blood or RNA data to the model did not improve the AUC.

The model AUC was increased considerably by adding characteristics of the intervention ('Precipitants') and perioperative changes in laboratory parameters ('Blood periop.') to the clinical data, and the highest overall AUC was achieved by a model using these three blocks of data (Fig. 4; Supplementary Figs 9 and 10). Adding transcript data to the model did not further improve AUC, but a model exploiting only preoperative and postoperative RNA data ('RNA+RNA periop.') showed almost identical performance.

The perioperative changes in mRNA abundance were more often predictive of POD than preoperative abundance in the 'RNA+RNA periop.' model. Most important transcripts were *BTN3A1* (butyrophilin), *LAP3* (leucine aminopeptidase 3), *DSN1* (DSN1 component of MIS12 kinetochore complex), *HPGD* (15-hydroxyprostaglandin dehydrogenase), and *KIF4B* (kinesin family member 4B). Notably, both preoperative *JAK2* (janus kinase) and circular *JAK2* mRNA were predictive of POD (Supplementary Fig. 11 and Supplementary Table 17). In an exploratory Cox regression analysis, we found that a considerable number of transcripts (*HPGD*, *BTN3A1*, *LAP3*, *JAK2*, and circular *JAK2*) were also associated with postoperative mortality (Supplementary Fig. 12 and Supplementary Table 18).

Discussion

We estimated POD prediction algorithms based on prospectively collected clinical, neuropsychological, blood-based, and neuroimaging data. This represents an early evaluation of non-routine data, including neuroimaging and gene expression, to enhance POD prediction from clinical variables through data-driven analyses.

By aggregating clinical preoperative data, precipitating factors with preoperative laboratory values, and postoperative changes, the model achieved good discriminability (AUC 0.83) with good model fit.

Previous approaches used retrospectively collected data or merged heterogenous data from multiple studies.[7,9,11] The only prospective study (SAGES) achieved an AUC of 0.71

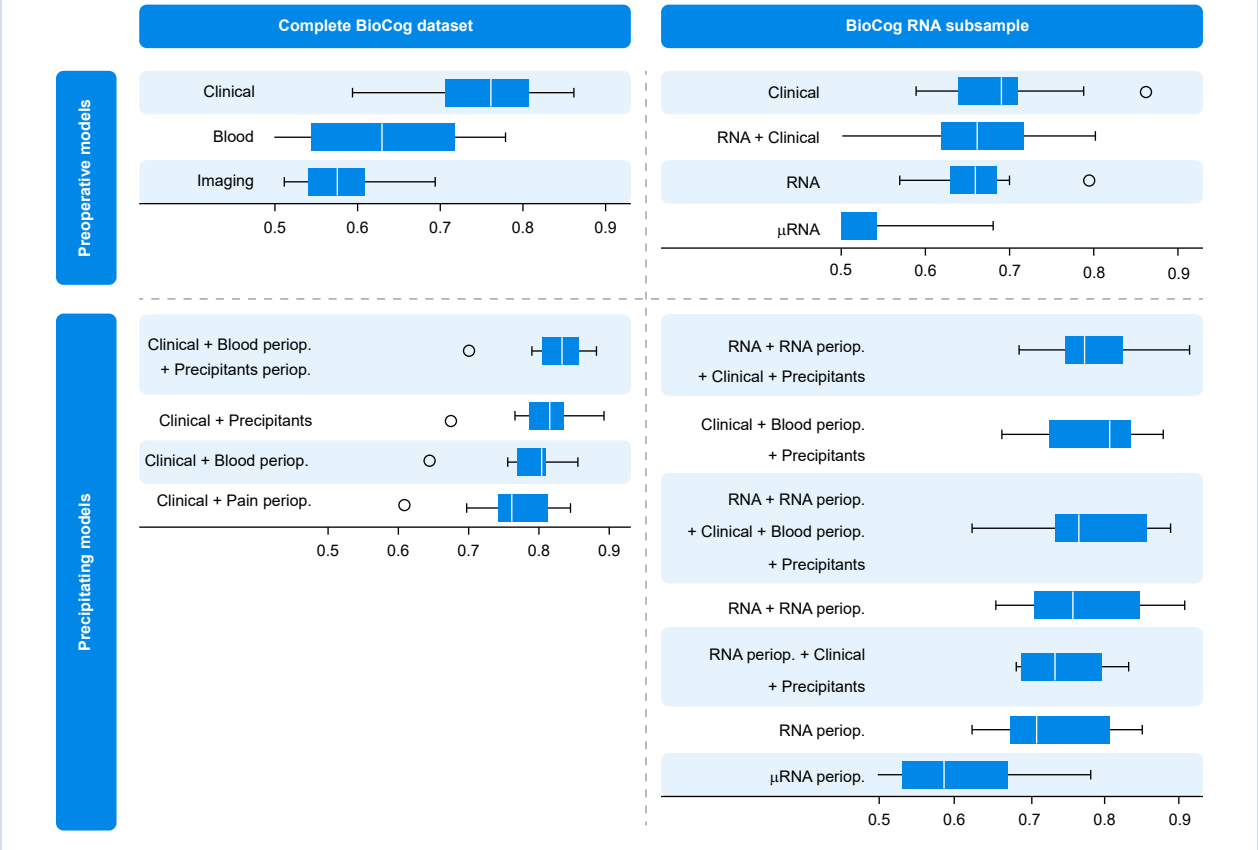

Fig 3. Boxplot displaying area under the curve (AUC) of the receiver-operating characteristic (ROC) over different folds of the cross-validation. A value of 1 indicates 100% sensitivity at 100% specificity, whereas a value of 0.5 indicates indiscriminability of the model for postoperative delirium (POD). Each model evaluates a different combination of available datasets, as indicated on the Y-axis. periop., perioperative (referring to precipitating factors, e.g. pain or medication, and perioperative changes in molecule abundance. Perioperative values for blood-based parameters including RNA were calculated as the difference between postoperative and preoperative values.); RNA, transcriptomic data features.

Table 2 Performance summary of GBT models. *Perioperative values for blood-based parameters including RNA were calculated as the difference between postoperative values and preoperative values. †Please note that GBT models inherently handle missing values, and hence, analyses can be conducted within the whole cohort, even for special data blocks such as 'imaging' with large portions of missing values. This allows comparison of performance parameters of models built from different data blocks as all refer to the same cohort and reflects real-word conditions where certain data may not be available for a certain patient (e.g. missing neuroimaging data owing to contraindications for MRI or urgent surgery). At each decision node, the algorithm learns an optimal default direction for samples with missing values by evaluating which branch yields the highest information gain. ‡Although GBT models can handle missing values inherently, models using RNA data were evaluated separately in complete-case analyses as transcript abundance was only available for a subgroup of patients. In practice, it can be assumed that mRNA data could be obtained from every patient, and therefore, we deemed predictive performance of gene expression data in a complete-case analysis more relevant. ¶These models were calculated for the subsample of patients with RNA data to compare model performance in this subgroup of patients. AUC, area under the curve; CI, confidence interval; GBT, gradient-boosted trees.

GBT model (combination of data blocks)	N†	AUC		Brier score	
		Mean	95% CI	Mean	95% CI
Models using exclusively preoperative data for prediction in the whole BioCog cohort					
Clinical	929	0.76	(0.69–0.81)	0.14	(0.13–0.16)
Clinical + blood	929	0.73	(0.68–0.79)	0.15	(0.14–0.15)
Blood	929	0.61	(0.54–0.68)	0.18	(0.16–0.20)
Imaging	929	0.58	(0.54–0.62)	0.23	(0.22–0.24)
Models using preoperative data and precipitating factors in the whole BioCog cohort					
Clinical + precipitants + blood periop.*	929	0.83	(0.79–0.86)	0.12	(0.12–0.13)
Clinical + precipitants	929	0.80	(0.77–0.84)	0.13	(0.12–0.14)
Clinical + blood periop.*	929	0.79	(0.75–0.82)	0.13	(0.13–0.14)
Clinical + pain	929	0.76	(0.72–0.80)	0.14	(0.13–0.15)
Blood periop.*	929	0.74	(0.68–0.77)	0.18	(0.16–0.20)
Precipitants	929	0.71	(0.68–0.75)	0.14	(0.14–0.15)
Pain	929	0.58	(0.56–0.6)	0.17	(0.16–0.18)
Models using exclusively preoperative data for prediction in the BioCog RNA subsample‡					
Clinical (for comparison‡)	371	0.69	(0.65–0.74)	0.16	(0.16–0.18)
RNA + clinical	371	0.66	(0.61–0.71)	0.17	(0.16–0.18)
RNA	371	0.66	(0.62–0.70)	0.17	(0.16–0.17)
µRNA	371	0.47	(0.40–0.54)	0.20	(0.19–0.22)
Model using preoperative data and precipitating factors in the BioCog RNA subsample‡					
Clinical + precipitants + blood periop.* (for comparison¶)	371	0.78	(0.73–0.83)	0.15	(0.14–0.16)
Clinical + precipitants + blood periop.* + RNA + RNA periop.*	371	0.77	(0.72–0.83)	0.15	(0.14–0.16)
Clinical + precipitants + RNA periop.*	371	0.74	(0.71–0.78)	0.16	(0.16–0.17)
RNA + RNA periop.*	371	0.77	(0.72–0.82)	0.15	(0.14–0.16)
RNA periop.*	371	0.73	(0.69–0.78)	0.16	(0.15–0.17)
µRNA periop.*	371	0.60	(0.54–0.66)	0.20	(0.19–0.21)

using machine learning in preoperative clinical data.[10] Our model solely relying on preoperative clinical data achieved similar performance (AUC 0.76). Notably, the SAGES study included patients with specific surgical procedures (joint replacement, spinal surgery, vascular surgery, and colectomy) and a narrower age range. Considering that age and procedure characteristics are important predictor variables in our models, our analyses may have benefitted from the larger variance in age and more diverse set of included surgical procedures in our cohort.

In our analyses, no improvement by adding preoperative non-routine data was achieved. Hence, thorough preoperative clinical evaluation to identify patients at risk can be considered a suitable approach in clinical routine. However, using algorithms as a diagnostic expert device can support quantifying POD risk and drive the establishment of POD risk assessment in routine clinical practice.

Results suggest that information about intervention and postoperative course can improve the model to an AUC of 0.8. Although models using precipitating data are intended for risk monitoring rather than prediction, relevant information is usually available before surgery, (i.e. estimated duration of intervention and expected postoperative pain), and may be used for prediction as well.

The prediction algorithm is intended for use by healthcare professionals and will be made available as a commercial

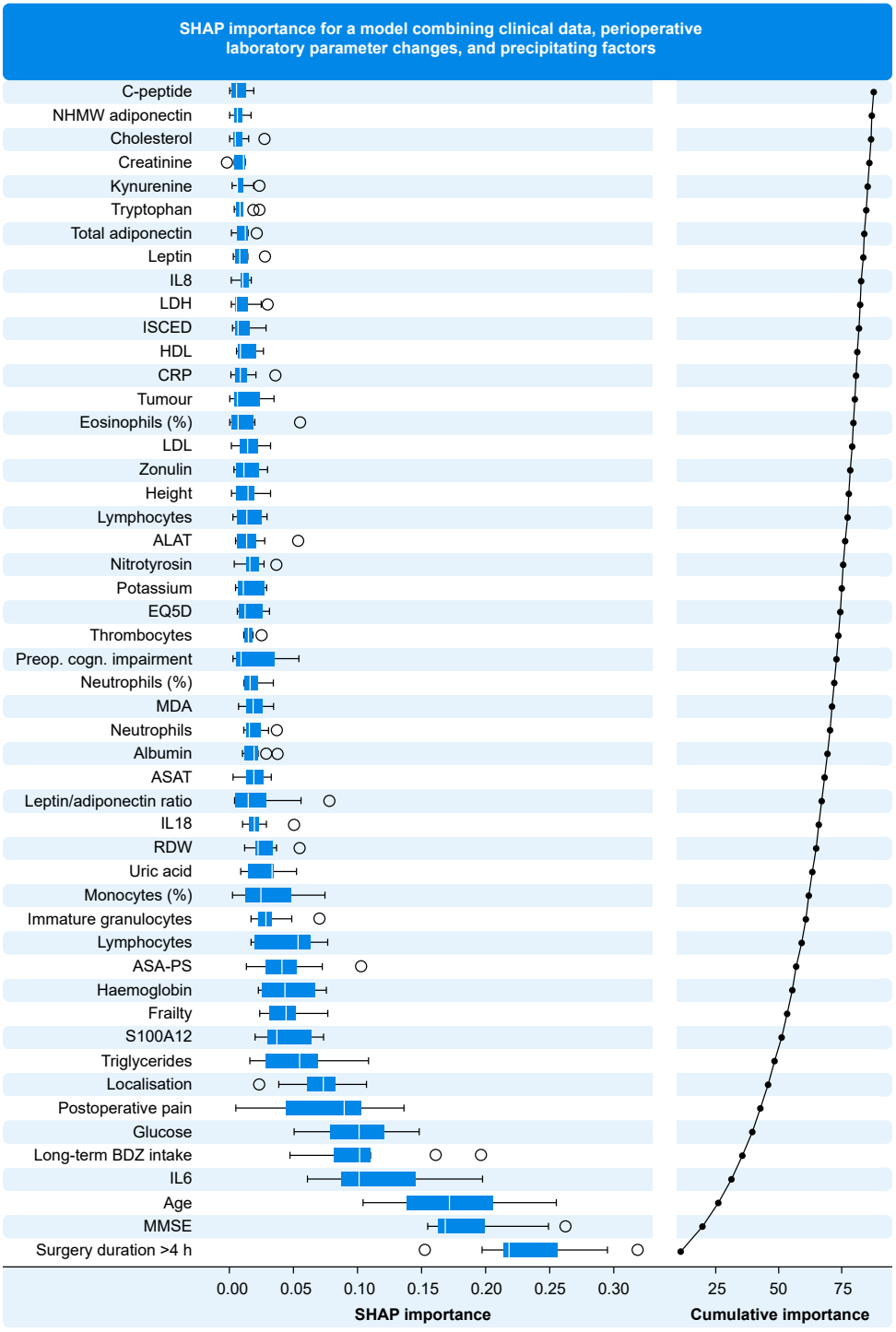

Fig 4. Feature importance of the model with the highest predictive performance. All blood-based parameters referred to in this figure, such as IL6, glucose, and triglycerides, refer to the perioperative change in laboratory parameters which had been calculated as the difference between postoperative and preoperative values. Tumour diagnosis includes solid malignancies, lymphoma, and leukaemia. ALAT, alanine aminotransferase; ASAT, aspartate aminotransferase; BDZ, benzodiazepine; CRP, C-reactive protein; HDL, high-density lipoprotein; IL, interleukin; ISCED, International Standard Classification of Education; LDH, lactate dehydrogenase; LDL, low-density lipoprotein; MDA, malondialdehyde; MMSE, Mini-Mental Status Examination; NHWM, non–high molecular weight; postop., postoperative; preop. cogn. impairment, preoperative cognitive impairment; RDW, red cell distribution width.

software. To assure practicability, GBT models were chosen which deal inherently with missing values in case of incomplete assessments.

Our analyses suggest that precipitating factors and perioperative laboratory assessments can considerably improve POD risk monitoring. Gene expression data may be of particular interest, as in the subgroup of patients with RNA data, a model exploiting only mRNA achieved an AUC similar to the best-performing model. A perioperative risk monitoring algorithm based on two gene expression analyses could relieve staff from extensive clinical assessments and be more cost-effective than using assays from multiple independent laboratories.

Many of the most predictive transcripts were mRNA of ubiquitously expressed genes involved in major molecular mechanisms such as cell proliferation (e.g. *DSN1*, *LAP3*, *KIF4B*, and *JAK*). This suggests that POD is a heterogeneous phenomenon originating via distinct molecular pathways. These central molecular nodes are nevertheless suitable for prediction. Certain transcripts suggest involvement of neuroinflammation and neuroplasticity (*BTN3A1*),[19] metabolic dysregulation and autophagy (*LAP3*),[20] proliferation (*DSN1*, *KIF4B*),[21] interaction with the immune system (*JAK*),[22,23] and senescence (*HPGD*).[24] Abovementioned molecules S100A12,[25] interleukins, and zonulin[26] point to certain immune response pathways, which may be related to neurotransmitter imbalance by tryptophan and kynurenine metabolism.[27] Some of the identified molecular targets have already been discussed with respect to neurodegeneration (i.e. malondialdehyde, nitrotyrosine,[28] metabolites of the kynurenine pathway,[29] S100A12,[25] and zonulin[26]).

BioCog was conducted in two study centres and therefore affected by centre bias induced by different language versions of questionnaires and screening list, preanalytical sample treatment (i.e. sample transportation to Berlin), MRI facilities, on-site laboratory procedures, and differences in health care systems. The BioCog study is small in relation to the wide spectrum of parameters included in our database. To fully exploit the potential of machine learning and to refine our models with subgroup analyses, larger samples are necessary, ideally from large real data pools of electronic patient charts. For instance, the current sample excluded patients without dementia with an MMSE score $\leq$23. However, in patients with progressive dementia, the configuration of parameters may differ to some extent, an even higher POD risk prediction accuracy might be also achieved, and brain atrophy may become a relevant biomarker. Conversely, integration of neuroimaging data into the machine learning models may have been complicated by both the smaller sample size and high collinearity between regional volume measures, which may boost overfitting and subsequently affect generalisability in models exploiting neuroimaging data. However, whether a single predictor derived from the neuroimaging dataset could improve the overall algorithm is a matter of future research. As external validation in an independent dataset is pending, we have used nested cross-validation as an internal validation procedure. The focus of this manuscript is prediction, whereas a molecular causal model cannot be addressed here. Single-variable analyses have not been adjusted for confounders, and so do our findings on γ-GT and transaminase changes in POD highlight the need for more comprehensive analyses that account for potential confounders and the data distribution: Elevated preoperative γ-GT in patients with POD could be confounded by preoperative cholestasis in those undergoing extensive upper gastrointestinal surgery (e.g. pancreatectomy or hemihepatectomy) with a major postoperative inflammatory response. After removal of the obstruction, γ-GT concentrations normalise, which may result in a pronounced postoperative decrease, but without true clinical relevance. In contrast, the postoperative increase in transaminases seen in patients with POD may reflect hepatic injury, possibly owing to intraoperative hepatic hypoperfusion, and could be causally linked to POD, independent of the surgical procedure. The best model was chosen by ROC-AUC, which is a measure of discrimination in diagnostic testing at the individual level. For prognostic questions at the individual patient level, reliable outcome probability estimation is essential. We therefore extended our analysis beyond discrimination (AUC) to formally assess model calibration. The low Brier score, supported by visual inspection of calibration and binned residual plots (Supplementary Fig. 10), supports the accuracy of the model over the whole range of predicted probabilities, making them suitable for clinical interpretation and demonstrating a strong model fit.[30]

POD screening was performed according to the evidence-based standard that measures POD at least twice a day and has a comprehensive geriatric assessment included to describe the clinical entity of this population. The clinical phenomenology was structured and annotated according to this standard.[2]

BioCog has made advancements towards POD prediction and will also facilitate comprehensive hypothesis-driven studies of patients for better understanding of pathophysiological processes and conception of interventional studies. Our dataset can guide prevention strategies to reduce POD (e.g. via the JAK pathway).[23]

Authors' contributions

Conceptualisation: FB, TBD, IF, DH, SK, DKM, PN, MP, TP, JP, AJCS, EAS, SW, CS, GW
Validation: FLL, FB, DH, MH, IMJK, SJTVM, CS
Data curation: FB, IF, DH, SH, MH, JJ, JKS, TP, CVH, JW, CS, GW
Software: SH, JW, A Wiehe, CS, GW
Visualisation: FLL, A Wiehe
Resources: LA, DKM, TP, HDV, CS, GW
Writing – original draft: FLL, MF, TP, CS, GW
Writing – review and editing: DB, FB, DKM, JP, PHS, AJCS, EAS, HDV
Methodology: LA, DB, JDB, NMDL, TBD, SG, IF, AF, JG, DH, SHH, JH, PH, IMJK, RK, JKS, SK, GL, CL, RM, AM, HJM, MN, PN, MP, SKP, TP, JP, RS, AJCS, EAS, EVD, SW, JW, A Wiehe, A Wolf, NZ, CS, GW
Formal analysis: DB, FB, JDB, SC, MMC, NMDL, TBD, SG, IF, IFDS, MF, AF, JG, SHH, JH, PH, AK, RK, JKS, SK, CL, HJM, MN, SKP, JP, RS, JW, A Wiehe, A Wolf, NZ
Investigation: FLL, LA, FB, JDB, SC, NMDL, TBD, SG, IFDS, MF, AF, JG, DH, FHY, SHH, MH, JH, PH, IMJK, RK, SK, GL, ML, CL, RM, AM, HJM, MN, KO, JP, RS, EAS, SJTVM, EVD, JMW, NZ
Supervision: TBD, JG, RK, SK, DKM, PN, TP, JP, EAS, CS, GW
Funding acquisition: TBD, PN, MP, TP, AJCS, EAS, SW, CS, GW
Project administration: PN, TP, JP, K Scheurer, K Scholtz, AJCS, EAS, CVH, A Wolf, CS, GW
Verification of the underlying data: FLL, GW

Acknowledgements

We thank the Koordinierungszentrum für Klinische Studien (KKS Berlin), and especially Alexander Krannich, and

ALTA (Siena, Italy) who provided additional administrative and coordinating services throughout the study. Multimodal data management was conducted by Pharmaimage Biomarker Solutions GmbH. The central biobank was provided by the Molecular Epidemiology Group, Max-Delbrück Center (MDC), Berlin, Germany. This group distributed samples for additional analyses to Atlas Biolabs GmbH, Immundiagnostik AG in Bernsheim, Germany, Institute of Protein Biochemistry at Consiglio Nazionale delle Ricerche di Pisa, Immune Study Lab of Institute of Medical Immunology and BIH Center for Regenerative therapies at Charité-Universitätsmedizin Berlin. We thank our colleagues who participated as MD students and team of study nurses, especially Tuba Gülmez, Emmanuel Keller, Mario Lamping, Helene Michler, Juliane Dörfler, Zdravka Bosancic, Fieras Nosierat, Irene Mergele, Victoria Windmann, and Anna Nottbrock. The authors further wish to thank the team of the student apprentices/interns of the Department of Anesthesiology at the Charité-Universitätsmedizin Berlin. Magnet resonance imaging has been supported by theBerlin Center for Advanced Neuroimaging core staff, especially Stefan Hetzer and Christian Labadie. Henning Krampe supported the study by recruiting and supervising students for neuropsychological testing.

Declarations of interest

GW (Pharmaimage Biomarker Solutions GmbH) is currently licensing a Class IIa and IVD (In Vitro Diagnostic) medical device: web-based machine learning software tool including laboratory diagnostics (mRNA signature) for pre- and perioperative risk prediction of POD, POCD, and postoperative mortality in clinical practice (patent application pending: EP26153633.8). An independent external validation study is currently ongoing at a separate clinical site to assess generalisability and real-world performance. GW is CEO of Pharmaimage Biomarker Solutions GmbH Berlin (Germany) and President of its subsidiary Pharmaimage Biomarkers Incl. (Cambridge, MA, USA). CS, GW, DB, TBD, SK, PN, TP, MP, AJCS, EAS, and SW report grants from the European Commission during the conduct of the study. Dr. Winterer reports grants from the Deutsche Forschungsgemeinschaft (DFG)/German Research Society and from the German Ministry of Health. CS reports grants from DFG/German Research Society, Einstein Foundation Berlin, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)/German Aerospace Center, Projektträger im DLR/Project Management Agency, Gemeinsamer Bundesausschuss (GBA)/Federal Joint Committee, inner university grants, Stifterverband/Non-Profit Society Promoting Science and Education, European Society of Anesthesiology and Intensive Care, BMWI – Federal Ministry of Economic Affairs and Climate Action, Dr. F. Köhler Chemie GmbH, Sintetica GmbH, Max-Planck-Gesellschaft zur Förderung der Wissenschaft e.V., Medtronic, BMBF – Federal Ministry of Education and Research, Robert Koch Institute and payments by Georg Thieme Verlag, board activity for Prothor, Takeda Pharmaceutical Company Ltd., Lynx Health Science GmbH, AWMF (Association of the Scientific Medical Societies in Germany), DFG, Deutsche Akademie der Naturforscher Leopoldina e.V. (German National Academy of Sciences Leopoldina), Berliner Medizinische Gesellschaft, European Society of Intensive Care Medicine (ESICM), European Society of Anaesthesiology and Intensive Care (ESAIC), Deutsche Gesellschaft für Anästhesiologie und Intensivmedizin (DGAI)/German Society of Anaesthesiology and Intensive Care Medicine, German Interdisciplinary Association for Intensive Care and Emergency Medicine (DIVI) as well as patents 15753 627.7, PCT/EP 2015/067731, 3 174 588, 10 2014 215 211.9, 10 2018 114 364.8, 10 2018 110 275.5, 50 2015 010 534.8, 50 2015 010 347.7, and 10 2014 215 212.7. GL and MH report grants from the BIH Charité Clinician Scientist Program during conduct of the study. TBD reports personal fees from Immundiagnostik AG during the conduct of the study. FLL and A. Wiehe report personal fees from Pharmaimage Biomarker Solutions GmbH during the conduct of the study. GL reports personal fees from Sobi, the University of Zurich and Thieme outside the submitted work. A. Wolf receives fees from the Kompetenz-Centrum Qualitätssicherung. EAS reports funding from Stephen Erskine Fellowship from Queens' College of the University of Cambridge, UK outside the BioCog study. JDB reports funding from Alzheimer Nederland outside of the study. JG received funding from the German Research Foundation (DFG), Federal Ministry of Education and Research (BMBF) and received payment for five lectures and presentations with about 1.500 € per presentation sponsored by Lundbeck, Janssen-Cilag, and Boehringer. SHH receives personal fees from Life&Brain GmbH. The remaining authors declare that they have no conflict of interest.

Funding

The research leading to these results has received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement n° 602461.

Data availability statement

Owing to the protection of intellectual property, machine learning algorithms will not be made publicly available, but can be obtained from GW (georg.winterer@pi-pharmaimage.com) after signing a confidentiality agreement. Participant data may be made available upon request after publication to researchers who provide a methodologically sound proposal in accordance with applicable legal and regulatory restrictions after careful review of each individual request. Access will only be granted in cases where the potential receiver of the data and purpose of the analysis is covered by the patients' informed consent and applicable legal regulations. Proposals for data analysis must be directed to both claudia.spies@charite.de and georg.winterer@pi-pharmaimage.com. Analyses will be limited to those approved in appropriate ethics and governance arrangements. All study documents that do not identify individuals (e.g. study protocol, informed consent form template) will be freely available on request.

Statement on sex and gender equity in research

Data on the patients' gender have not been collected, and we refer to 'sex' throughout the manuscript. Although the best-performing model was found to have similar discriminability in women and men, analyses in the Supplementary material have neither been adjusted nor disaggregated for sex, although further in-depth analyses would require this step to yield valid results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2026.01.025.

References

1. Androsova G, Krause R, Winterer G, Schneider R. Biomarkers of postoperative delirium and cognitive dysfunction. *Front Aging Neurosci* 2015; **7**: 112
2. Aldecoa C, Bettelli G, Bilotta F, 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. *Eur J Anaesthesiol* 2024; **41**: 81–108
3. Aldecoa C, Bettelli G, Bilotta F, et al. European Society of Anaesthesiology evidence-based and consensus-based guideline on postoperative delirium. *Eur J Anaesthesiol* 2017; **34**: 192–214
4. Winterer G, Androsova G, Bender O, et al. Personalized risk prediction of postoperative cognitive impairment - rationale for the EU-funded BioCog project. *Eur Psychiat* 2018; **50**: 34–9
5. Yang T, Yang H, Liu Y, et al. Postoperative delirium prediction after cardiac surgery using machine learning models. *Comput Biol Med* 2024; **169**, 107818
6. Zhao XX, Li JL, Xie XH, et al. Online interpretable dynamic prediction models for postoperative delirium after cardiac surgery under cardiopulmonary bypass developed based on machine learning algorithms: a retrospective cohort study. *J Psychosom Res* 2024; **176**, 111553
7. Oosterhoff JHF, Karhade AV, Oberai T, Franco-Garcia E, Doornberg JN, Schwab JH. Prediction of postoperative delirium in geriatric hip fracture patients: a clinical prediction model using machine learning algorithms. *Geriatr Orthop Surg Rehabil* 2021; **12**, 21514593211062277
8. Rossler J, Shah K, Medellin S, et al. Development and validation of delirium prediction models for noncardiac surgery patients. *J Clin Anesth* 2024; **93**, 111319
9. Bishara A, Chiu C, Whitlock EL, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol* 2022; **22**: 8
10. Racine AM, Tommet D, D'Aquila ML, et al. Machine learning to develop and internally validate a predictive model for post-operative delirium in a prospective, observational clinical cohort study of older surgical patients. *J Gen Intern Med* 2021; **36**: 265–73
11. Dodsworth BT, Reeve K, Falco L, et al. Development and validation of an international preoperative risk assessment model for postoperative delirium. *Age Ageing* 2023; **52**, afad086
12. Saczynski JS, Kosar CM, Xu G, et al. A tale of two methods: chart and interview methods for identifying delirium. *J Am Geriatr Soc* 2014; **62**: 518–24
13. Feinkohl I, Borchers F, Burkhardt S, et al. Stability of neuropsychological test performance in older adults serving as normative controls for a study on postoperative cognitive dysfunction. *BMC Res Notes* 2020; **13**: 55
14. German National Cohort (GNC) Consortium. The German National Cohort: aims, study design and organization. *Eur J Epidemiol* 2014; **29**: 371–82
15. Harrell F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. New York: Springer; 2015
16. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. *J Gerontol A Biol Sci Med Sci* 2001; **56**: M146–56
17. Kaiser MJ, Bauer JM, Ramsch C, et al. Validation of the Mini Nutritional Assessment short-form (MNA-SF): a practical tool for identification of nutritional status. *J Nutr Health Aging* 2009; **13**: 782–8
18. Gaudreau JD, Gagnon P, Harel F, Tremblay A, Roy MA. Fast, systematic, and continuous delirium assessment in hospitalized patients: the nursing delirium screening scale. *J Pain Symptom Manage* 2005; **29**: 368–75
19. Ribot JC, Lopes N, Silva-Santos B. gammadelta T cells in tissue physiology and surveillance. *Nat Rev Immunol* 2021; **21**: 221–32
20. Feng L, Chen Y, Xu K, et al. Cholesterol-induced leucine aminopeptidase 3 (LAP3) upregulation inhibits cell autophagy in pathogenesis of NAFLD. *Aging (Albany NY)* 2022; **14**: 3259–75
21. Zhu C, Zhao J, Bibikova M, et al. Functional analysis of human microtubule-based motor proteins, the kinesins and dyneins, in mitosis/cytokinesis using RNA interference. *Mol Biol Cell* 2005; **16**: 3187–99
22. Xue C, Yao Q, Gu X, et al. Evolving cognition of the JAK-STAT signaling pathway: autoimmune disorders and cancer. *Signal Transduct Target Ther* 2023; **8**: 204
23. Rodriguez S, Hug C, Todorov P, et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat Commun* 2021; **12**: 1033
24. Sun CC, Zhou ZQ, Yang D, et al. Recent advances in studies of 15-PGDH as a key enzyme for the degradation of prostaglandins. *Int Immunopharmacol* 2021; **101**, 108176
25. Lai Y, Lin P, Lin F, et al. Identification of immune microenvironment subtypes and signature genes for Alzheimer's disease diagnosis and risk prediction based on explainable machine learning. *Front Immunol* 2022; **13**, 1046410
26. Boschetti E, Caio G, Cervellati C, et al. Serum zonulin levels are increased in Alzheimer's disease but not in vascular dementia. *Aging Clin Exp Res* 2023; **35**: 1835–43
27. Romer TB, Jeppesen R, Christensen RHB, Benros ME. Biomarkers in the cerebrospinal fluid of patients with psychotic disorders compared to healthy controls: a systematic review and meta-analysis. *Mol Psychiatry* 2023; **28**: 2277–90
28. Jomova K, Vondrakova D, Lawson M, Valko M. Metals, oxidative stress and neurodegenerative disorders. *Mol Cell Biochem* 2010; **345**: 91–104
29. Giil LM, Midttun O, Refsum H, et al. Kynurenine pathway metabolites in Alzheimer's disease. *J Alzheimers Dis* 2017; **60**: 495–504
30. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–38

Handling Editor: Robert Sanders