BJA Open

*BJA Open*, 8 (C): 100236 (2023)

**doi:** 10.1016/j.bjao.2023.100236
Advance Access Publication Date: 2 November 2023
Original Research Article

ORIGINAL RESEARCH ARTICLE

# Validation of a convolutional neural network that reliably identifies electromyographic compound motor action potentials following train-of-four stimulation: an algorithm development experimental study

Richard H. Epstein[1,*], Olivia F. Perez[1], Ira S. Hofer[2], J Ross Renew[3], Réka Nemes[4] and Sorin J. Brull[3]

[1]Department of Anesthesiology, Pain Management and Perioperative Medicine, University of Miami Miller School of Medicine, Miami, FL, USA, [2]Department of Anesthesiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [3]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic College of Medicine and Science, Jacksonville, FL, USA and [4]Department of Anesthesiology and Intensive Care, University of Debrecen, Debrecen, Hungary

*Corresponding author. Department of Anesthesiology, Pain Management and Perioperative Medicine, 1400 NW 12th Ave, Suite 4022F, Miami, FL, USA 33136. E-mail: repstein@med.miami.edu

Prior presentations: Part of this work has been accepted for presentation as an abstract at the 2023 Annual Meeting of the American Society of Anesthesiologists, San Francisco, CA, USA.

## Abstract

**Background:** International guidelines recommend quantitative neuromuscular monitoring when administering neuromuscular blocking agents. The train-of-four count is important for determining the depth of block and appropriate reversal agents and doses. However, identifying valid compound motor action potentials (cMAPs) during surgery can be challenging because of low-amplitude signals and an inability to observe motor responses. A convolutional neural network (CNN) to classify cMAPs as valid or not might improve the accuracy of such determinations.
**Methods:** We modified a high-accuracy CNN originally developed to identify handwritten numbers. For training, we used digitised electromyograph waveforms (TetraGraph) from a previous study of 29 patients and tuned the model parameters using leave-one-out cross-validation. External validation used a dataset of 19 patients from another study with the same neuromuscular block monitor but with different patient, surgical, and protocol characteristics. All patients underwent ulnar nerve stimulation at the wrist and the surface electromyogram was recorded from the adductor pollicis muscle.
**Results:** The tuned CNN performed highly on the validation dataset, with an accuracy of 0.9997 (99% confidence interval 0.9994–0.9999) and $F_1$ score=0.9998. Performance was equally good for classifying the four individual responses in the train-of-four sequence. The calibration plot showed excellent agreement between the predicted probabilities and the actual prevalence of valid cMAPs. Ten-fold cross-validation using all data showed similar high performance.
**Conclusions:** The CNN distinguished valid cMAPs from artifacts after ulnar nerve stimulation at the wrist with >99.5% accuracy. Incorporation of such a process within quantitative electromyographic neuromuscular block monitors is feasible.

**Keywords:** electromyography; machine learning; neural network; neuromuscular block; train-of-four

The American Society of Anesthesiologists and the European Society of Anaesthesiology and Intensive Care recently published guidelines strongly recommending quantitative neuromuscular monitoring when using neuromuscular blocking agents.[1,2] Assessment of the train-of-four count (TOFC) is a critical component of such monitoring because it informs the choice and dose of reversal agent (sugammadex *vs* neostigmine),[1,2] provides information related to the depth

**Received:** 17 August 2023; **Accepted:** 3 October 2023

of paralysis (e.g. return of the second twitch in the train-of-four [TOF] corresponds to 93% first twitch (T1) depression from baseline for atracurium and vecuronium[3]), and allows assessment and maintenance of deep levels of neuromuscular block using the post-tetanic count (PTC).[4]

To evaluate the TOFC or the PTC, a neuromuscular monitor must determine if a valid compound motor action potential (cMAP) has occurred after nerve stimulation (Fig 1) and differentiate it from noise.

While, conceptually, a determination that an elicited motor response represents a cMAP is simple, such is not the case in clinical practice. At deep levels of neuromuscular block, the signal-to-noise ratio is small, often making such determinations difficult because electrical noise may be misinterpreted as a cMAP or the valid cMAP waveform may be altered by such interference. A common complaint from anaesthesia practitioners new to quantitative monitoring of neuromuscular function is 'The TOFC displayed by the monitor was 0, yet the patient was breathing or moving'. Such a lack of convergent validity between clinical observation and a monitor's output can reduce confidence in the device's accuracy and forestall acceptance of quantitative monitoring into routine clinical practice. These apparent discrepancies between subjective and objective assessments increase when the baseline twitch amplitude is low because of factors such as the amount of subcutaneous fat[5] or improper stimulating or recording electrode placement (e.g. not directly over the ulnar nerve or the corresponding muscle being monitored). Furthermore, the TOFC may differ if assessed by visual inspection, palpation, mechanomyography, accelerometry, or electromyography (EMG).[6 7] Manufacturers' algorithms to identify valid cMAP responses are proprietary and can vary between versions of the same device. Some changes (e.g. modification of the threshold amplitude required to identify a twitch) can result in calculation of a different TOFC from the same set of electromyogram signals in the TOF. Furthermore, tucking a patient's arms under the surgical drapes during laparoscopic, robotic, or neurosurgical procedures will adversely affect monitor accuracy when using methods that require unencumbered thumb movement (e.g. accelerometry, kinemyography), and interfere with subjective TOFC detection. Finally, improper placement of stimulating electrodes or high impedance from inadequate skin preparation can make the TOFC displayed by the monitor inaccurate.

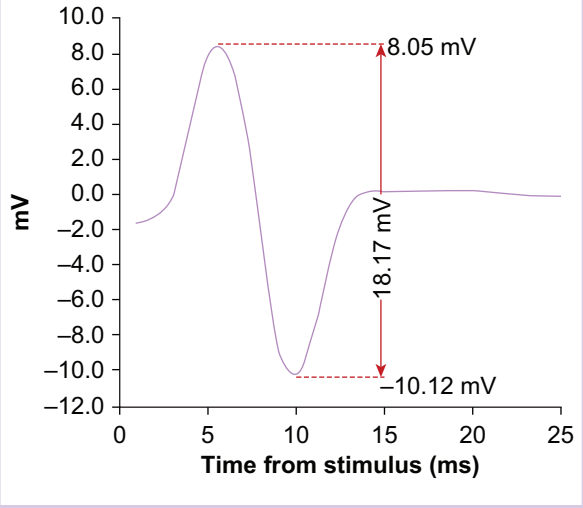


Fig. 1. Archetypal compound motor action potential (cMAP) from a representative study subject before administration of the neuromuscular blocking agent. This trace displays the cMAP from the adductor pollicis muscle as recorded from surface electrodes after stimulation of the ulnar nerve at the volar wrist at time 0. The initial positive deflection of the recording represents the sum of all activated motor unit depolarisations and the subsequent trough, the sum of all motor unit repolarisations. The amplitude of the response is the difference between the peak and trough of the signal. Recordings continued for 100 ms after the stimulus, but as there is no meaningful information related to motor activity beyond 20 ms, the analysed signals were truncated at 20 ms. Traces are frequently not as pristine as shown but are often altered by superimposed noise or shifted from the nominal baseline of 0 mV. A range of amplitudes and varying peak and trough locations occur among patients.

Because prescribed therapy may be influenced substantively by technology and patient considerations, exploring alternative and potentially less ambiguous methods to measure the TOFC and PTC is worthwhile. In a clinical setting, it is not possible to directly measure neuromuscular block of the diaphragm; rather, one must make inferences based on peripheral nerve stimulation (e.g. the ulnar nerve) and motor responses (e.g. from the adductor pollicis muscle). Ensuring the absence of diaphragmatic movement requires intense levels of neuromuscular block such that only several twitches after post-tetanic stimulation are present.[8] Because monitors often systematically undercount valid cMAP responses at deep levels of block, titrating neuromuscular blocking agents to achieve that goal can be challenging.

When using EMG, the cMAP has a typical shape, varying within individual patients at different levels of neuromuscular block primarily by a scaling factor (amplitude) rather than morphology (Fig 1). The shape of a cMAP resembles the letter z when rotated ~70° clockwise, with variation among patients. Thus, by analogy, cMAP classification by a convolutional neural network (CNN) is feasible because this machine learning method works extremely well for identifying handwritten digits and letters, whose morphology also varies.[9] We hypothesised that a published CNN used to identify digits would perform at least as well for recognising valid cMAP responses after TOF stimulation.

## Methods

The institutional review board of the University of Miami determined on 16 March 2023 that this analysis of deidentified EMG waveform data obtained from prior studies that had obtained informed consent does not constitute human subjects research. The randomised clinical trials from which the data were provided were approved by an institutional review board (Mayo Clinic, Jacksonville, FL, #20–000629) or an ethical board (University of Debrecen, Debrecen, Hungary, No. OGYÉI2690/2018), and informed patient consent was obtained from all enrolled subjects. The 2015 version of the Standards for Reporting Diagnostic accuracy studies (STARD) checklist was followed.[10]

### Data sources

Data used for training the CNN were provided by one of the co-authors (JRR) from 29 adult patients enrolled at Mayo Clinic,

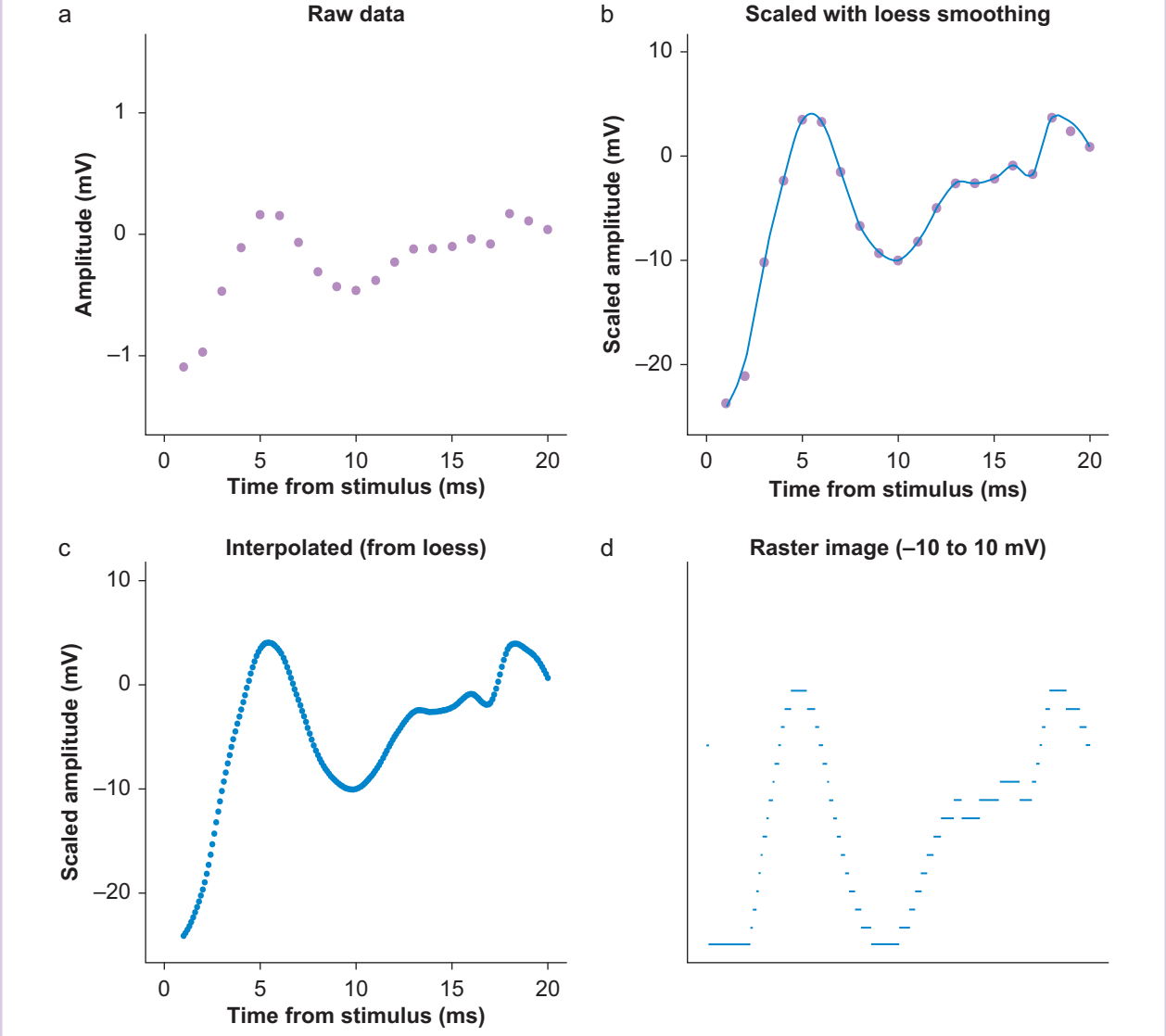


Fig. 2. Sequential preprocessing of the digitised compound motor action potential (cMAP). A low-amplitude cMAP from the adductor pollicis muscle after delivery of a single 50-mA stimulus lasting 200 μs to the ulnar nerve at the wrist is shown. Panel (a) plots the raw data from a partially paralysed patient who received a dose of rocuronium. The recording interval was every 1 ms for 100 ms, with the first 20 ms after the stimulus used for classifying the response as a valid cMAP. The amplitude (peak minus trough, measured in mV at 3–8 ms [0.16 mV] and 8–13 ms [−0.46 mV], respectively) was 0.62 mV. In panel (b), the voltages were scaled by 21.74 so that the maximum of the absolute value of the peak and trough values was 10 mV (i.e. 21.74=10.0/0.46). Then, a smooth curve was constructed using the locally estimated scatterplot smoothing (LOESS) method with span=0.25. In panel (c), the LOESS line was used to interpolate the voltages from 1.0 to 20.0 ms at 0.1 ms intervals. In panel (d), the values from panel (c) were converted into a raster image with of 191 × 201 pixels (38 391 pixels per image). Values <−10 mV were set equal to −10 mV, and those above 10 mV were set equal to 10 mV. The image was then used as the input to the convolutional neural network (see Fig 3). The process resulted in each cMAP being scaled to approximately the same size. Each cMAP was manually tagged as valid or invalid using an Excel workbook (Microsoft, Redmond, WA, USA) with a visual inspection of the smoothed plots of the voltage *vs* time data. For each patient, the cMAP was considered to represent a valid response if the shape of the curve followed the contour of the baseline cMAP and the peak and trough locations were close to those obtained from the baseline for the patient.

Jacksonville, as part of a study in which the TetraGraph EMG (Senzime AB, Uppsala, Sweden) neuromuscular block monitor was used.[11] In that investigation, patients' arms were tucked under surgical drapes for elective robotic or laparoscopic surgery. All patients included in the current study received rocuronium for neuromuscular relaxation, mostly inhalation anaesthesia for maintenance, deep neuromuscular block throughout surgery, and sugammadex for antagonism of neuromuscular block.[11]

The data used for external validation of the CNN were provided by another co-author (RN) from 19 American Society of Anesthesiologists physical status 1–3 patients aged ≥18 yr

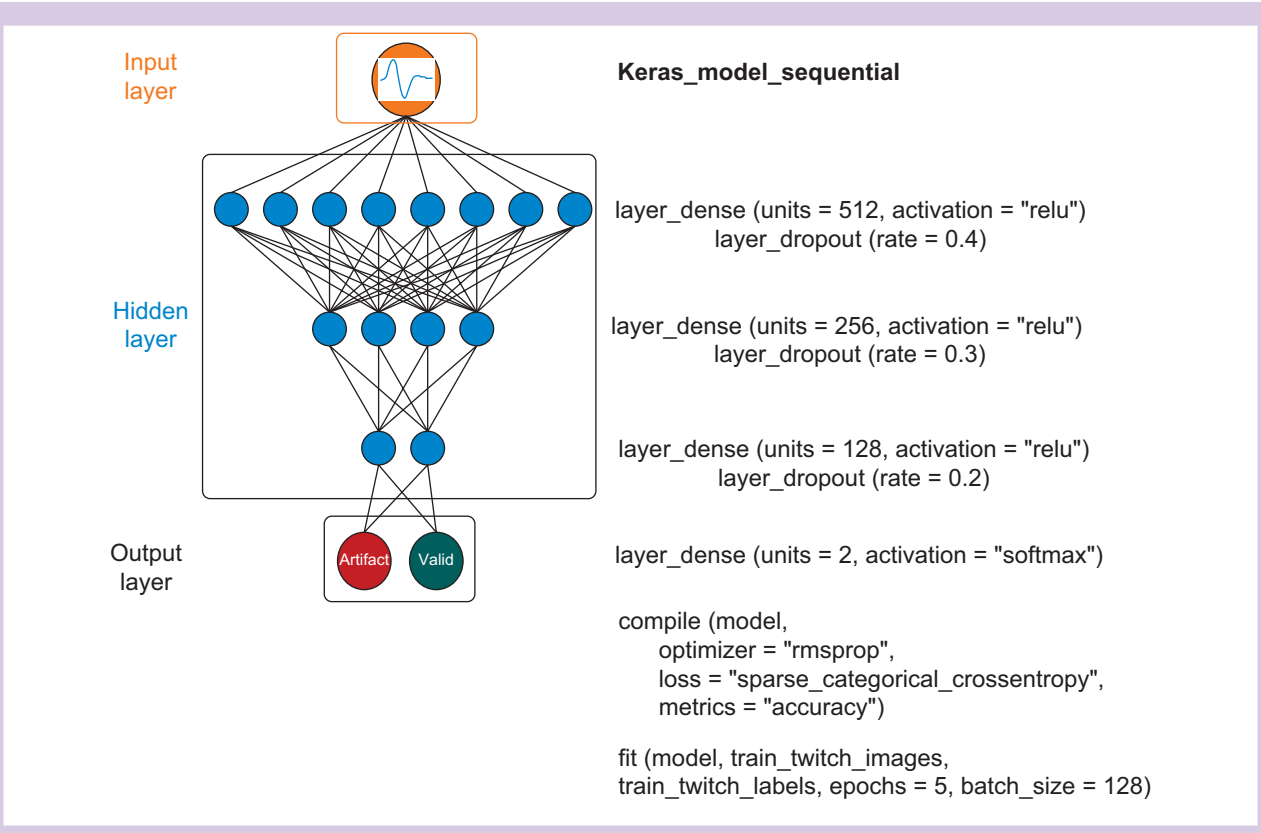


Fig. 3. Schematic description of the initial ('out-of-box') convolutional neural network (CNN) used for the binary classification of a valid compound motor action potential (cMAP) or a non-response. This CNN was modified from a published algorithm for classifying handwritten digits or characters in the Modified National Institute of Standards and Technology (MNIST) dataset. The CNN used for this study consists of a single input, three hidden layers, and two outputs. The input is the raster image of the processed EMG waveform at the adductor pollicis or abductor digiti minimi muscles after electrical stimulation of the ulnar nerve, as described in Fig 1. The three sequential hidden layers have 512, 256, and 128 nodes with rectified linear unit (relu) activation and a dropout applied after each to reduce overfitting. The model was fit using five epochs with a batch size of 128. The output layer uses softmax activation to assign a probability that the waveform is a valid cMAP or a non-response.

undergoing elective surgery given rocuronium for neuromuscular block. Anaesthesia was maintained with a target-controlled propofol infusion, and the TetraGraph was used to monitor the depth of neuromuscular block.[12] The protocol was designed to facilitate spontaneous recovery to a TOF ratio (TOFR) >0.9, with neostigmine administered if this endpoint was not reached. Spontaneous recovery produced a wider range of cMAP amplitudes that increased more slowly and more smoothly than in the training dataset. Thus, a more extensive evaluation of the CNN could be performed.

In both studies, anonymised data files containing the raw, digitised EMG waveforms had been uploaded to a secure server (TetraConnect, Senzime AB) maintained by the manufacturer and made available for analysis. These files consisted of surface EMG voltages recorded over the adductor pollicis muscle from 1 to 100 ms after stimulation of the ulnar nerve at the wrist. The current and pulse width necessary to achieve supramaximal stimulation were determined automatically by the monitor for each patient and ranged from 40 to 60 mA and 200–300 μs, respectively. For each pulse, the stimulation mode (TOF or single twitch), the elapsed time from the start of data recording, and the amplitude of the peak and trough of each cMAP were provided. Also, the file included the stimulus current (mA) and pulse width (μs) and the relevant output from each TOF neurostimulation sequence (i.e. the TOFR, TOFC, or PTC).

Sufficient details are provided in the remaining sections of the Methods to allow other researchers to replicate our process. For readers mostly interested in the Results, these sections can be skimmed or disregarded.

## Data preprocessing

Preprocessing and tagging were required to prepare the raw EMG data for input to the CNN. The individual waveform data were uploaded into an Excel workbook (Microsoft, Redmond, WA, USA) and code written in Visual Basic for Applications (Microsoft) to facilitate visual exploration and manual classification of the electromyographic response to each stimulus of the ulnar nerve. The cMAP, corresponding to depolarisation and repolarisation of the motor unit, was completed within 18 ms after ulnar nerve stimulation, with peaks nearly always occurring between 3 and 8 ms and troughs between 8 and 13 ms after stimulation (Fig 1). Thus, the EMG data were truncated at 20 ms because the signal after this interval was

uninformative. Two authors (RHE, OFP) independently classified each twitch, autoscaled by Excel in the y-axis to produce a uniformly sized plot based on the maximum and minimum voltage recorded as a binary value (valid cMAP or not). The waveform was classified as representing a valid cMAP if the morphology resembled that of the archetypal twitch (Fig 1) and the location of the peak and trough (i.e. the elapsed time in ms after the stimulus) matched the locations of the baseline twitch. Valid cMAPs with amplitudes as small as 0.05 mV were often discernible with this approach, but no arbitrary amplitude cut-off was applied. The first author (RHE) re-examined all waveforms for which there was a discrepancy between the evaluators or where the other evaluator (OFP) was unsure, then assigned a final classification. All twitches (single twitch or part of a TOF) were included in the training dataset. Only the responses elicited from TOF stimulation were included in the external validation dataset, with each component identified as the first through fourth twitch (T1, T2, T3, and T4).

After tagging, each 20-ms portion of the EMG waveform was processed (Fig 2) in RStudio 2022.12.0 (RStudio, Boston, MA, USA) running R version 4.2.1 (R Foundation, Vienna, Austria). To summarise the process described in detail in the legend of Fig 2, data were scaled such that the maximum of the absolute value of the peak or trough voltage was 10 mV, a smooth line was fit through the data points, and then the predicted voltages (truncated to 10 mV or −10 mV if outside that range) were determined at 0.1 ms intervals between 1 and 20 ms. Then, each curve was converted to a raster image 191 pixels wide × 201 pixels high for processing by the CNN. Rasterisation is required because CNNs are designed to deal with images, not time-series amplitude data.

Because the immense size of the image tensor resulted in long execution times during training, and occasional insufficient memory exceptions when the combined data were processed, we evaluated the impact on CNN performance of reducing the image resolution. In parallel, we decreased the time resolution from 0.1 to 0.2 and 0.25 ms and the voltage resolution from 0.1 to 0.2 and 0.25 mV during cross-validation of the combined datasets. This data reduction step lowered the image size from 191 × 201 pixels (38 391 pixels) to 96 × 101 pixels (9696 pixels) and 77 × 81 pixels (6237 pixels), respectively, reducing the computation times.

To further evaluate the performance of the CNN, the two datasets were combined from the 46 patients and $n$=47 016 twitches elicited after TOF stimulation (i.e. 11 754 TOF stimulations). The amplitude of each twitch was then calculated for each waveform as the difference between the peak voltage (occurring between 2.8 and 8.2 ms after the stimulus) and the trough voltage (occurring between 6.4 and 12.6 ms after the stimulus). The original classifications were then rechecked by printing the waveforms to a PDF file after digital scaling and interpolation, with each set of twitches from a TOF on a single row and four sets per page to facilitate comparisons of the peaks and troughs. Each of the 47 016 waveforms was re-evaluated independently by RHE and SJB (based on morphology and the interval from stimulus to the peak and trough) and classified as valid or not. Disagreements were resolved by subsequent collaborative reinspection of the waveforms by these investigators. There were 527 reassignments made (1.12%), with 59.2% representing changes to a valid twitch from an invalid twitch and 40.8% from a valid to an invalid twitch.

### Neural network development

A CNN in RStudio was built based on a published example of code used to process the Modified National Institute of Standards and Technology (MNIST) image dataset of handwritten digits (Fig 3).[13 14] The packages required included *ggplot2*, *dplyr*, *tidyverse*, *keras*, *reticulate*, *tensorflow*, and *caret*. The input to the neural network was the raster image corresponding to the cMAP, as described in the previous section. A dropout layer was added to the example code after each dense layer in the three hidden layers of the CNN to reduce overfitting. Instead of 10 outputs in the MNIST example, corresponding to digits 0 to 9, there were two outputs, corresponding to a valid cMAP or absence of a response.

The CNN was applied to the training dataset using the original 191 × 201 pixel images, and performance was assessed for the 29 subjects by 28 leave-one-out cross-validation runs. Among the 28 runs, the means and 99% confidence intervals (CIs) were calculated for the sensitivity, specificity, positive predictive value, negative predictive value, precision, recall, $F_1$ Score, and accuracy of the classifier.

The model hyperparameters were tuned for the training dataset, using the 96 × 101 pixel images and leave-one-out cross-validation based on consideration of performance criteria, memory usage, and execution time. A limited grid search was performed varying the number of nodes in the three hidden layers from (512, 256, 128) to (256, 128, 64), the number of epochs from 2, 3, or 4, and the batch size from 8, 16, or 32. Thus, there were 2 × 3 × 3=18 distinct models, each cross-validated 28 times with one different subject held out from among the 29 subjects for each validation run. Performance metrics were calculated among the 28 cross-validation results for each model. The tuned model was then used to assess the performance and calibration of the external validation dataset. In addition, the distribution of correct classifications by the CNN as related to the amplitudes of the manually identified valid twitches was determined. This process incorporated three types of external validation: temporal (different study dates), population (different patient groups and neuromuscular blocking agent administration protocol), and geographic (from different institutions).[15]

The performance impact of reducing the image size from 191 × 201 pixels to 96 × 101 pixels and then 77 × 81 pixels was assessed by 10-fold cross-validation of the combined training and external validation datasets. This analysis was performed because computational times and memory requirements are substantively affected by image size.

As a final evaluation of the performance of the CNN, the model using the tuned hyperparameters was applied to the combined dataset (with the additional cycle of manual validation, as described in the last paragraph of the previous section) and performance was evaluated with 10-fold cross-validation (five patients held out of the 46 total patients in each fold).

## Results

There were 28 025 cMAP responses evaluated among 29 patients in the training dataset and 20 912 among 19 patients in the external validation dataset. There were 5228 individual responses for each twitch (T1-T4) in the TOF in the validation dataset. Substantive differences were present in the two study populations, an important feature of an external validation process (Table 1).[15] In the training dataset, case durations were considerably longer (mean 3.94 *vs* 1.20 h of monitoring), and

there was a much higher percentage of absent twitches recorded at the adductor pollicis in response to ulnar nerve stimulation (77.6% *vs* 18.1%). These differences reflected the study protocols (laparoscopic/robotic surgery with deep paralysis *vs* all elective surgery with spontaneous neuromuscular recovery). In addition, the average frequency of monitoring was much lower in the training dataset (approximately one TOF sequence per minute) than in the external validation dataset (approximately four TOF sequences per minute). Amplitudes were not normally distributed in either the training or external validation datasets. Baseline T1 amplitudes were higher in the training dataset (median 13.9 mV *vs* 10.6 mV, $P$=0.022 by the Wilcoxon rank-sum test, Table 1). Baseline amplitudes ranged from 2.8 mV to 28.7 mV, Table 1). Among the 48 037 cMAPs (T1-T4), amplitudes were higher in the external validation dataset (median 4.1 mV *vs* 2.4 mV, $P$<0.0001, Table 1).

The unmodified CNN[13] performed well on cross-validation of the training dataset, with an overall accuracy of 0.987 (99% CI 0.974–1.000) and $F_1$ score=0.977 (99% CI 0.974–1.000, Table 2). CNN performance improved after hyperparameter tuning (Table 2). The optimal hyperparameters were three hidden layers with 512, 256, and 128 nodes (no change), respectively, four epochs (reduced from five), and a batch size of 32 (reduced from 128). From the leave-one-out cross-validation, results were: overall accuracy=0.999 (99% CI 0.997–1.00), precision=1.000 (99% CI 1.000–1.000), recall=0.999 (99% CI 0.997–1.000), and $F_1$ score=0.999 (99% CI 0.999–1.000).

Image size reduction from 191 × 201 pixels to 96 × 101 pixels resulted in equivalent overall model performance (accuracy 1.000 [99% CI 1.000–1.000] *vs* 0.992 [99% CI 0.974–1.000]; $F_1$ score 1.000 [99% CI 0.999–1.000] *vs* 0.993 [99% CI 0.977–1.000] (Supplementary Table S1). However, further reduction to 77 x 81 pixels reduced performance substantively (accuracy 0.929 [99% CI 0.876 to 0.982]; F1 score 0.890 [99% CI 0.803 to 0.977], Supplementary Table S1). Thus, the 96 × 101 pixel images were used for external validation of the CNN.

The tuned CNN performed more highly than the raw CNN on the external dataset among all twitches, with an accuracy=0.9997 (99% CI 0.9994–0.9999) and $F_1$ score=0.9998 (Table 3). Performance for these metrics was equally good for classifying each of the four individual responses (T1-T4) in the TOF sequence (Table 3).

The calibration plot for the external validation dataset showed excellent agreement between the predicted probabilities and the actual prevalence of valid twitches, with all data points lying on the line of identity (Supplementary Fig. S1).

The tuned CNN had a similar high performance on the combined dataset as evaluated by 10-fold cross-validation (five random patients from the 46 withheld, without replacement, for each fold), with accuracy=0.998 (95% CI 0.995–1.001) and $F_1$ score=0.997 (95% CI 0.994–1.001) (Table 2). Among the 18 962 twitches classified manually as valid, the CNN correctly identified 99.75%. There were only two twitches that were assessed as invalid manually and the CNN provided classification as a valid cMAP. The CNN was able to reliably detect cMAPs classified manually as valid with amplitudes even as low as 0.05 mV.

## Discussion

The CNN described had excellent performance for classifying valid cMAPs from the adductor pollicis muscle after electrical stimulation of the ulnar nerve. After hyperparameter tuning, model accuracy on an external validation dataset was at least as good as that of the best-reported average accuracies of neural network algorithms for handwritten digit recognition (99.3%–99.5%).[9] This comparison is relevant because digit and character recognition have been used commercially for many years with a high degree of success in a variety of real-world processes (e.g. recognition of zip codes by automated mail sorting machines, license plate scanners used by law enforcement, mobile bank deposits from cell phone cameras). The excellent correlation between the predicted and actual probabilities of identifying a valid cMAP from the calibration plot (Supplementary Fig. S1) and the ability to detect cMAPs with extremely low amplitudes (Fig 2) further demonstrates the utility of our method to determine validly the TOFC or the PTC. Nonetheless, achieving a strong EMG signal is important for the reliable use of quantitative neuromuscular block monitors currently in clinical use. We stress the importance of carefully prepping the skin, applying the sensor in the proper location, allowing for adequate curing time of the silver/silver-chloride electrodes once placed (up to 10 min), and measuring the baseline twitch as soon as the patient loses consciousness, before administering the neuromuscular blocking agent.[16]

It is likely that each electromyographic neuromuscular block monitor model would need to have a dedicated CNN developed to identify valid cMAPs because signal processing algorithms applied for artifact rejection vary among manufacturers. Signal processing and filtering alters the morphology of the underlying waveform, for example, the electrocardiogram when one switches from the standard monitoring mode to a diagnostic mode (e.g. to allow viewing of

Table 1 Dataset characteristics. IQR, inter-quartile range; mV, millivolts; SD, standard deviation.

| | Dataset | |
|---|---|---|
| | Training | External validation |
| Number of subjects | 29 | 19 |
| Hours of monitoring, mean (SD) | 3.94 (1.28) | 1.20 (0.33) |
| Responses analysed per case, mean (SD) | 966 (770) | 1101 (317) |
| Responses analysed, *n* | 28 025 | 20 912 |
| Valid, *n* (%) | 6273 (22.4) | 17124 (81.9) |
| Non-response, *n* (%) | 21 752 (77.6) | 3788 (18.1) |
| All valid amplitudes, mV, median (IQR) | 2.4 (1.0–6.1) | 4.1 (1.4–7.0) |
| Baseline amplitude, mV, median (IQR, range) | 13.9 (11.0–18.0, 2.9–28.7) | 10.6 (7.5–14.6, 3.1–19.1) |

Table 2 Performance of the convolutional neural network among the $n$=28 leave-one-out cross-validation runs from the training dataset, and 10-fold cross-validation for the combined dataset. CI, confidence interval; cMAP, compound motor action potential; $F_1$ score, the harmonic mean of the precision and recall. [a] Model with original hyperparameters as used for handwritten digit recognition.[10]. Hidden layers=3 (512, 256, 128 nodes); image resolution 191 × 201 pixels (0.1 ms and 0.1 mV); epochs=5; batch size=128. [b] Model with hyperparameter tuning based on leave-one-out cross-validation. Hidden layers=3 (512, 256, 128 nodes); image resolution 96 × 101 pixels (0.2 ms and 0.2 mV); epochs=4; batch size=32. [c] Combined training and testing dataset, tested using the original hyperparameters noted in first table footnote but with the image resolution in second table footnote because of memory constraints. There were 47 016 images evaluated, 53.1% of which were classified manually as valid cMAPs and 46.9% as non-responses (i.e. not valid cMAPs). In each fold, five patients were withheld for testing. The detection prevalence in the holdout samples ranged from 33.1% to 74.0%.

| | Leave-one-out cross-validation | | 10-Fold cross-validation |
|---|---|---|---|
| Performance metric | "Out-of-box" model[a] Mean (99% CI) | Tuned model[b] Mean (99% CI) | Combined data[c] Mean (99% CI) |
| Sensitivity | 0.959 (0.921–0.997) | 0.999 (0.997–1.000) | 0.995 (0.988–1.002) |
| Specificity | 1.000 (0.999–1.000) | 1.000 (1.000–1.000) | 1.000 (1.000–1.000) |
| Positive predictive value | 0.999 (0.998–1.000) | 1.000 (1.000–1.000) | 1.000 (0.999–1.000) |
| Negative predictive value | 0.981 (0.963–1.000) | 0.999 (0.997–1.000) | 0.995 (0.989–1.001) |
| Precision | 0.999 (0.998–1.000) | 1.000 (1.000–1.000) | 1.000 (0.999–1.000) |
| Recall | 0.959 (0.921–0.997) | 0.999 (0.997–1.000) | 0.995 (0.988–1.002) |
| $F_1$ score | 0.977 (0.954–1.000) | 0.999 (0.999–1.000) | 0.997 (0.994–1.001) |
| Accuracy | 0.987 (0.974–1.000) | 1.000 (0.999–1.000) | 0.998 (0.995–1.001) |

pacer spikes). As a corollary, revalidation of the CNN would need to be performed if modifications to digital or hardware filters between model versions substantively altered waveforms. CNN model programming is straightforward, but the manual effort to tag a sufficient number of traces is considerable. Typically, CNNs require at least several thousand instances of each class for training; for example, the MNIST digit dataset consists of 60 000 training and 10 000 testing images among the 10 digits.[13] However, incorporation within a quantitative EMG neuromuscular block monitor is much less resource intensive and would be achievable within the typical computer systems and programming languages embedded in such devices. For example, the TetraGraph software is written in C++, which is well-suited for neural network image processing.[17] The embedding of a CNN to classify cMAPs into clinically used EMG monitors is the ultimate objective of this research and would have a substantively positive impact on the quality of the analytics performed by those devices.

Our study identified valid cMAPs for each patient during the tagging process based on morphology and correspondence of the peak and trough to their baseline location, where the signal was always robust. We did not have any information as to whether a palpable or visible motor contraction of the thumb was present. However, we question the validity of such subjective assessments. We know from clinical estimates of the TOFR that anaesthesia practitioners cannot reliably differentiate a 50% difference in strength between the first and fourth twitch after TOF stimulation either when assessed visually or tactilely.[18,19] Inferentially, because the forces generated by the adductor pollicis muscle at deep levels of neuromuscular block are much smaller (e.g. <5% of baseline) it is unclear if counting such twitches would be accurate. Moreover, the hand is often inaccessible for visual or tactile assessment of the response to TOF stimulation during surgery.

Correlation between quantitative and qualitative assessment of the TOFC depends on the baseline first twitch cMAP amplitude because the degree of muscle paralysis is inversely related to the ratio of the first twitch amplitude to the baseline (control) first twitch amplitude (T1/T1c). This ratio is the relevant metric when determining the onset time of neuromuscular block (i.e. 95% first twitch depression) and clinical duration (i.e. time to 25% spontaneous recovery).[20] Thus, there will be a greater degree of neuromuscular block for the same first twitch amplitude when the baseline amplitude is high

Table 3 External validation of the tuned convolutional neural network. CI, confidence interval; $F_1$ Score, the harmonic mean of precision and recall.

| | Train-of-four twitch analysed (N) | | | | |
|---|---|---|---|---|---|
| Performance metric | All (20 912) | T1 (5228) | T2 (5228) | T3 (5228) | T4 (5228) |
| Sensitivity | 0.9996 | 0.9994 | 0.9995 | 0.9997 | 1.0000 |
| Specificity | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Positive predictive value | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Negative predictive value | 0.9984 | 0.9935 | 0.9976 | 0.9992 | 1.0000 |
| Precision | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Recall | 0.9996 | 0.9994 | 0.9995 | 0.9997 | 1.0000 |
| $F_1$ score | 0.9998 | 0.9997 | 0.9998 | 0.9999 | 1.0000 |
| Accuracy (95% CI) | 0.9997 (0.9994–0.9999) | 0.9994 (0.9983–0.9999) | 0.9996 (0.9986–1.0000) | 0.9998 (0.9989–1.0000) | 1.0000 (0.9993–1.0000) |

than when it is low. For example, an amplitude of 1.0 mV reflects 80% first twitch depression when the baseline is 5 mV but 95% depression when the baseline is 20 mV. This dichotomy strongly suggests that counting twitches to determine block onset or assessing when additional doses of relaxant are needed will be highly variable among patients, given the considerable range in baseline amplitudes observed clinically. Rather, a better strategy would be for the quantitative monitor to identify first that there is a valid first twitch in the TOF and then, if true, compute the ratio of the amplitude of that twitch to the baseline (obtained at supramaximal current) value (T1/T1c) and display the magnitude of twitch depression. This normalisation provides independence from the baseline amplitude and eliminates the need to apply an arbitrary threshold value. Feedback controllers that adjust infusions of neuromuscular blocking drugs to maintain a constant level of neuromuscular block are based on T1 depression, not the TOFC.[21,22] Providing the magnitude of T1 depression (compared with baseline T1) would facilitate delivery of infusions to maintain moderate to deep levels of block.

We note that our proposed process of identifying valid cMAPs using a CNN is most relevant to managing moderate to deep levels of neuromuscular block,[23] when the TOFC is only 1 or 2. Nonetheless, the algorithm works extremely well when the signal amplitude is large. For assessing that the TOFR is >0.9 to confirm adequate recovery of neuromuscular function before extubation, routine engineering signal processing algorithms, as currently applied, are sufficient because the T1 will have returned to close to its baseline value of at least several mV. At the time of extubation, the clinical issue related to assessing adequate recovery of neuromuscular function is that practitioners cannot accurately assess the TOFR visually or tactilely, despite being able to correctly discern the number of twitches present (TOFC).

The approach we followed likely can be generalised to categorising other waveforms with typical morphologies commonly encountered during routine anaesthesia and intensive care practice, such as those from invasive pressure monitors, photoplethysmographs, capnographs, and thermodilution cardiac output devices. Essentially, these share the characteristic of a waveform that ascends and then descends after some periodic event (e.g. cardiac contraction, exhalation, saline injection). One only needs to convert the waveform to a raster image, and then the CNN will process the relevant visual information to generate the most probable output class. Thus, we anticipate that the approach we outline should work well for recognition of overdamped or underdamped arterial waveforms, bronchospasm, spontaneous breathing, or cardiac oscillations in the capnograph, and invalid thermodilution curves.

### Limitations and strengths

Only a single manufacturer's quantitative neuromuscular block monitor was analysed, so the weighting factors in the hidden layers of the CNN would likely not be optimal for every device. However, the approach described in this study can be followed, and a similarly high classification performance of the CNN would be expected because EMG waveforms (cMAPS) are fundamentally the same. A strength of the study is that we performed external validation and confirmed calibration of the CNN using a large dataset of waveforms from another institution obtained in a group of patients studied under a very different protocol. External validation and calibration are steps often omitted from machine learning assessments in medicine, with only internal validation performed.[24] Other populations in which the described CNN needs to be validated before clinical use include infants, children, and patients with underlying neuromuscular diseases. However, the output layer of the CNN can easily be modified from the current two classes (valid or invalid) to multiple classes (e.g. valid adult, valid paediatric, invalid) or extended to include alternative pathways (e.g. a different hidden layer based on patient age or presence of a specific neuromuscular disorder that affects peripheral nerve conduction or the neuromuscular junction).

The absence of information related to the presence of visible or tactile twitches corresponding to the digitised waveforms is a potential limitation related to implementation but is not relevant to the evaluation of the CNN's performance. The absence of clinical correlation is irrelevant to our study because our goal was to identify valid cMAP waveforms, not subjectively observed twitches. As we explain above, decision-making regarding the management of moderate to deep levels of neuromuscular block would be improved by considering the extent of first-twitch depression, not the TOFC. However, having a more reliable method to assess the TOFC during spontaneous recovery and to determine the PTC are also clinically valuable.

## Conclusions

We modified a convolutional neural network (CNN) developed originally for recognising handwritten digits to classify EMG waveforms recorded at the adductor pollicis as a valid compound motor action potential relevant to determining the train-of-four count. The CNN had extremely high accuracy on both the training dataset and an external dataset performed in a different geographical location, on a different group of patients undergoing different surgical procedures, and under a different protocol of muscle relaxation. The approach described should have utility in improving the determination of the train-of-four count responses, a measurement critical to selecting the appropriate drug and dose for reversing neuromuscular block and monitoring the level of paralysis during surgical cases.

## Authors' contributions

Conceptualisation: RHE
Validation: RHE, OFP, SJB
Software: RHE
Resources: RHE, JRR, RN, SJB
Data curation: RHE
Formal analysis: RHE, OFP
Writing original draft: RHE
Writing review and editing: all authors
Visualisation: RHE, ISH, SJB

## Declarations of interests

SJB has intellectual property assigned to Mayo Clinic (Rochester, MN); is a consultant for Merck & Co., Inc. (Kenilworth, NJ); is a principal, shareholder, and Chief Medical Officer in Senzime AB (Uppsala, Sweden); and an unpaid member of the scientific/clinical advisory boards for The Doctors Company (Napa, CA); Coala Life Inc. (Irvine, CA); NMD Pharma (Aarhus, Denmark); and Takeda Pharmaceuticals (Cambridge, MA). JRR has ongoing industry-sponsored research (Merck & Co., Inc.) with funds to his employer and has served on a scientific advisory board for Senzime AB (Uppsala, Sweden). ISR

is the founder and president of Extrico Health, an informatics company that helps hospitals leverage data from their electronic health record for decision-making purposes, receives research support and serves as a consultant for Merck, and is funded, in part, by National Institutes of Health (NIH) grant 1K01HL150318. The other authors declare that they have no conflicts of interest.

## Funding

This work was supported solely by Departmental sources.

## Acknowledgements

We acknowledge the assistance of Nicolas Alberti, Linux Systems Administrator at the University of Miami Center for Computation Science, who configured the university's GPU-accelerated high-performance supercomputer (https://idsc.miami.edu/Triton/) for use in some of the neural network analysis.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bjao.2023.100236.

## References

1. Thilen SR, Weigel WA, Todd MM, et al. American Society of Anesthesiologists practice guidelines for monitoring and antagonism of neuromuscular blockade: a report by the American Society of Anesthesiologists task force on neuromuscular blockade. *Anesthesiology* 2023; **138**: 13–41
2. Fuchs-Buder T, Romero CS, Lewald H, et al. Peri-operative management of neuromuscular blockade: a guideline from the European society of Anaesthesiology and intensive care. *Eur J Anaesthesiol* 2023; **40**: 82–94
3. O'Hara DA, Fragen RJ, Shanks CA. Reappearance of the train-of-four after neuromuscular blockade induced with tubocurarine, vecuronium or atracurium. *Br J Anaesth* 1986; **58**: 1296–9
4. Naguib M, Brull SJ, Johnson KB. Conceptual and technical insights into the basis of neuromuscular monitoring. *Anaesthesia* 2017; **72**(Suppl 1): 16–37
5. Kuiken TA, Lowery MM, Stoykov NS. The effect of subcutaneous fat on myoelectric signal amplitude and cross-talk. *Prosthet Orthot Int* 2003; **27**: 48–54
6. Bhananker SM, Treggiari MM, Sellers BA, Cain KC, Ramaiah R, Thilen SR. Comparison of train-of-four count by anesthesia providers versus TOF-Watch® SX: a prospective cohort study. *Can J Anaesth* 2015; **62**: 1089–96
7. Bowdle A, Bussey L, Michaelsen K, et al. Counting train-of-four twitch response: comparison of palpation to mechanomyography, acceleromyography, and electromyography. *Br J Anaesth* 2020; **124**: 712–7
8. Werba A, Klezl M, Schramm W, et al. The level of neuromuscular block needed to suppress diaphragmatic movement during tracheal suction in patients with raised intracranial pressure: a study with vecuronium and atracurium. *Anaesthesia* 1993; **48**: 301–3
9. Tabik S, Peralta D, Herrera-Poyatos A, Herrera F. A snapshot of image pre-processing for convolutional neural networks: case study of MNIST. *Int J Comput Intell Syst* 2017; **10**: 555–68
10. Bossuyt PM, Reitsma JB, Bruns DE, et al. *Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies*. https://www.equator-network.org/reporting-guidelines/stard/. [Accessed 5 July 2023]
11. Hernandez V, Chaves-Cardona H, Renew JR, Brull SJ. Electromyographic and acceleromyographic monitoring in restricted arm movement surgical setting: a prospective, randomized trial. In: *Annual Meeting of the American Society of Anesthesiologists*; October 9-13, 2021. A2069. San Diego, CA. (abstract) Anesthesiology
12. Nemes R, Lengyel S, Nagy G, et al. Ipsilateral and simultaneous comparison of responses from acceleromyography- and electromyography-based neuromuscular monitors. *Anesthesiology* 2021; **135**: 597–611
13. Chollet F, Kalinowski T, Allaire JJ. *Deep learning with R*. Manning; Shelter Island, New York. https://www.manning.com/books/deep-learning-with-r-second-edition. [Accessed 5 July 2023]
14. Deng L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process Mag* 2012; **29**: 141–2
15. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021; **14**: 49–58
16. Brull SJ, Silverman DG. Neuromuscular block monitoring. In: Ehrenwerth J, Eisenkraft JE, editors. *Anesthesia equipment: principles and Applications*. St. Louis: Mosby Year Book; 1993. p. 297–318
17. Wang Y-Q, Limare N. A fast C++ implementation of neural network backpropagation training algorithm: application to Bayesian optimal image demosaicing. *IPOL* 2015; **5**: 257–66
18. Viby-Mogensen J, Jensen NH, Engbaek J, Ørding H, Skovgaard LT, Chraemmer-Jørgensen B. Tactile and visual evaluation of the response to train-of-four nerve stimulation. *Anesthesiology* 1985; **63**: 440–2
19. Brull SJ, Silverman DG. Visual and tactile assessment of neuromuscular fade. *Anesth Analg* 1993; **77**: 352–5
20. Power SJ, Jones RM. Relationship between single twitch depression and train-of-four fade. *Anesth Analg* 1987; **66**: 633–6
21. Jaklitsch RR, Westenskow DR. A model-based self-adjusting two-phase controller for vecuronium-induced muscle relaxation during anesthesia. *IEEE Trans Biomed Eng* 1987; **34**: 583–94
22. Janda M, Simanski O, Bajorat J, Pohl B, Noeldge-Schomburg GF, Hofmockel R. Clinical evaluation of a simultaneous closed-loop anaesthesia control system for depth of anaesthesia and neuromuscular blockade. *Anaesthesia* 2011; **66**: 1112–20
23. Brull SJ, Kopman AF. Current status of neuromuscular reversal and monitoring: challenges and opportunities. *Anesthesiology* 2017; **126**: 173–90
24. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLOS Med* 2012; **9**, e1001221

*Handling editor: Phil Hopkins*