Novel Preoperative Risk Stratification Using Digital Phenotyping Applying a Scalable Machine-Learning Approach

Pascal Laferrière-Langlois[1,2,3,4 *], MD; Fergus Imrie[5], DPhil; Marc-Andre Geraldo[2,3], MSc; Theodora Wingert[1], MD; Nadia Lahrichi[2], PhD; Mihaela van der Schaar[6,7], PhD; Maxime Cannesson[1], MD, PhD.

---

1. Department of Anesthesiology and Perioperative Medicine, UCLA David Geffen School of Medicine, Los Angeles, USA **2.** Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Montreal, Quebec, Canada **3.** Maisonneuve-Rosemont Hospital Research Center, Montréal, Québec, Canada **4.** Department of Anesthesiology and Pain Medicine, Maisonneuve-Rosemont Hospital, CIUSSS de l'Est de L'Ile de Montréal, Montréal, Québec, Canada **5.** Department of Electrical and Computer Engineering, UCLA, Los Angeles, USA **6.** Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK **7.** The Alan Turing Institute, London, UK

***Corresponding Author:** Dr Pascal Laferrière-Langlois; Pascal.laferriere-langlois@umontreal.ca. 5415 Bld de l'Assomption, Montreal (QC), Canada, H1T 2M4; Phone: +1-514-252-3400

Prior Presentation: ASA conference 2022, Best of Abstracts session (October 23rd 2022) and poster session (October 22nd 2022)

Abbreviated Title: Digital phenotyping for surgical patients

Summary Statement: Digital phenotyping exploits machine learning to subdivide heterogeneous populations, like a pre-surgical population, into more homogeneous subgroups based on readily available digital data. In this context, digital phenotyping will enlighten the different typical profiles of patients undergoing each surgery. As a proof of concept, we showed that pre-surgical phenotypes exist for three frequently performed surgeries and, not only can they help describe these typical profiles, but they can be attributed pre-operatively to risk stratify the patients and predict adverse postoperative outcomes.

Funding and Acknowledgements: This work was supported by National Institute of Health (Bethesda,

Maryland) grant Nos. R01- R01EB029751, R01HL144692, National Science Foundation grant No 1722516, by the *Fond de recherche du Québec en Santé* No 322164 and by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Competing Interests: No competing is relevant to the work reported or conclusion. Dr Laferriere-Langlois has ownership interest in Divocco Medical (Montreal, Canada) and Divocco AI (Montreal, Canada). Dr. Cannesson, is a consultant for Edwards Lifesciences (Irvine, California) and has ownership interest in Perceptive Medical Inc. (Newport Beach, California), and in Sironis (Newport Beach, California). Dr Imrie, Dre Wingert, Pr Lahrichi and Pr Van der Schaar have no competing interest to declare.

Author Contribution: MC had the original idea. PLL, TW and MC provided clinical expertise, and first selected the features from the medical records. PLL and FI wrote the software and conducted the experiments. FI, MAG, NL and MVDS revised the software and confirmed the methodology. PLL and FI drafted the manuscript and all authors revised, and approved, the final manuscript.

Abbreviations:

AUROC: Area Under Receiver Operating Characteristics curve

ARI: Adjusted Rand index

ASA: American Society of Anesthesiologists

EHR: Electronic Health Record

ICD: International Classification of Disease

ICU: Intensive Care Unit

IQR: Interquartile Range

LVEF: Left Ventricular Ejection Fraction

LOS: Length of Stay

NMI: Normalized Mutual Info Score

PACU: Post-Anesthesia Care Unit

PDW: Patient Data Warehouse

POSPOM: PreOperative Score to Predict PostOperative Mortality

SD: Standard Deviation

UCLA: University of California in Los Angeles

Abstract

Introduction

Classification of perioperative risk is important for patient care, resource allocation, and guiding shared decision-making. Using discriminative features from the electronic health record (EHR), machine learning algorithms can create digital phenotypes among heterogenous populations, representing distinct patient subpopulations grouped by shared characteristics, from which we can personalize our care, anticipate clinical care trajectories, and explore therapies. We hypothesized that preoperative digital phenotypes exist in pre-operative settings and are associated with postoperative adverse events including in-hospital and 30-day mortality, 30-day surgical redo, intensive care unit (ICU) admission, and hospital length of stay (LOS).

Methods

We identified all laminectomies, colectomies, and thoracic surgeries performed over a 9-year period from a large hospital system. Seventy-seven readily extractable preoperative features were first selected from clinical consensus, including demographics, medical history, and lab results. Three surgery-specific datasets were built and split into derivation and validation cohorts using chronological occurrence. Consensus *k*-means clustering was performed independently on each derivation cohort, from which phenotypes' characteristics were explored. Cluster assignments were used to train a random forest model to assign patient phenotypes in validation cohorts. We reconducted descriptive analyses on validation cohorts to confirm characteristics similarities with derivation cohorts, and quantified the association of each phenotype with postoperative adverse events by using the area under *receiver operating characteristic curve* (AUROC). We compared our approach to ASA alone and investigated a combination of our phenotypes with the ASA score.

Results

A total of 7,251 patients met inclusion criteria, of which 2,480 were held out in a validation dataset based on chronological occurrence. Using segmentation metrics and clinical consensus, three distinct phenotypes were created for each surgery. The main features used for segmentation included urgency of the procedure, pre-operative LOS, age, and comorbidities. The most relevant characteristics varied for each of the three surgeries. Low-risk phenotype *alpha* was the most common (2039/2480, 82%) while high-risk phenotype *gamma* was the rarest (302/2480, 12%). Adverse outcomes progressively increased from phenotypes *alpha* to *gamma*, including 30-day mortality (0.3%, 2.1% and 6.0%, respectively), in-

hospital mortality (0.2%, 2.3% and 7.3%) and prolonged hospital LOS (3.4%, 22.1% and 25.8%). When combined with ASA score, digital phenotypes achieved higher AUROC than ASA score alone (hospital mortality: 0.91 vs. 0.84; prolonged hospitalization: 0.80 vs 0.71).

Conclusion

For three frequently performed surgeries, we identified three digital phenotypes. The typical profiles of each phenotype were described and could be used to anticipate adverse postoperative events.

Keywords

Machine learning; Digital phenotyping; Perioperative outcomes; Outcome prediction; Artificial intelligence

Key points summary

Question

Can we use the electronic health record to attribute a digital phenotype to pre-surgical patients undergoing laminectomy, colectomy, or thoracic surgery, and use this phenotype to better understand a patient's profile and anticipate their care trajectory?

Findings

For each of the three surgical cohorts investigated, we identified three surgical phenotypes with specific clinical characteristics, which can be used to predict adverse postoperative trajectories including mortality, prolonged hospital length of stay, admission to intensive care unit, and surgical reoperation.

Meaning

Without human inference, we can suggest pre-operative phenotypes that can inform care providers on the profile of patients undergoing surgery, whether further preoperative evaluation may be beneficial, potentially anticipate prolonged hospitalization, or to personalize the consent process.

Background

More than 240 million surgeries are performed globally each year and postoperative mortality, despite remaining below 2%, is described as the third leading cause of worldwide mortality [1, 2]. Patients considered to be in a high-risk surgical population will account for 80% of this mortality [3]. Preoperative risk is therefore critical in order to risk stratify patients in order to optimize resource allocation, conduct preoperative interventions, and share the decision-making between the patients and their providers [3].

Multiple attempts at preoperative risk stratification have been published. The POSPOM Score (PreOperative Score to Predict PostOperative Mortality)[4] and Charlson comorbidity index[5] are two scores focusing on mortality, while recent other risk stratification tools also aim to predict intermediate outcomes such as organ failure [6]. Nonetheless, the ASA score developed in 1941 by the American Society of Anesthesiologists (ASA), remains the most widely used score due to its simplicity and generalizability. Despite not being developed originally to predict complications, it has been shown to correlate with post-operative risks [7, 8]. However, the ASA score has several significant weaknesses: it disregards the type of surgery as a risk factor, relies on the anesthesiologist's experience, and an ASA score of 3 (intermediate) is overutilized [8-10]. Recently, machine learning (ML) algorithms have been applied to electronic health record (EHR) data and have demonstrated the potential to improve risk prediction[11-15]. However, to date, most studies have presented models with supervised learning trained to predict specific post-operative complications, including mortality, cardiorespiratory adverse events, allergic reaction, as well as the ASA score itself[16-18].

Digital phenotyping is a machine learning methodology that can be applied to heterogeneous populations to identify subgroups sharing common characteristics[19-21]. Phenotyping algorithms can identify discriminating features and discover homogeneous subgroups, uncovering patterns and commonalities

that may not be perceptible to the individual or by classical statistical techniques. The features identified from these types of algorithms can then be interpreted by clinicians[22]. By defining the digital phenotype of each population subgroup, one can then suggest future behaviours based on the behaviour of the other members in the subgroup [19, 23]. This strategy has been explored primarily in psychiatric and neurological conditions [24], but recent studies have expanded towards specific cohorts within perioperative medicine[25], such as patients undergoing joint arthroplasty[26, 27]

In this manuscript, we hypothesized that unsupervised ML algorithms, specifically consensus k-means clustering, can create digital phenotypes of preoperative patients who share common key pre-operative characteristics, and that these phenotypes are associated with postoperative complications, including 30-day mortality, in-hospital mortality, 30-day reoperation, ICU admission, and prolonged hospital length of stay (LOS) defined by LOS greater than the 90th percentile. To confirm the generalizability and scalability of our approach across the spectrum of surgeries, we selected three non-cardiac surgeries based on two criteria: 1) frequently performed surgeries in the United States, and 2) presenting a different perioperative risk profile to ensure that our phenotyping algorithm remains relevant in a wide spectrum of risk. We created surgery-specific phenotypes for laminectomy, colectomy, and thoracic surgery with thoracotomy or thoracoscopy.

Methods

This retrospective study was approved by the University of California in Los Angeles (UCLA) Institutional Review Board (UCLA-A IRB#15-000518); patients' written consent was waived due to the retrospective approach of this study. This research was conducted and reported in accordance with guidelines [28], which we adapted for phenotyping, an unsupervised task. Further details regarding the methods are included in the supplementary material.

Database

The data was extracted from the Patient Data Warehouse (PDW), a custom-built database described in a previously in detail, containing perioperative data from all surgeries completed within University of California Los Angeles (UCLA) Health System since the inception of EHR in March 2013[29]. To populate the PDW, the data is first extracted from the EPIC Clarity database (EPIC Systems, Madison, WI, USA) before being extracted and validated into the PDW, comprising more than 4,000 distinct perioperative features.

We extracted all surgical records that were performed between March 1st, 2013 and April 1st, 2022 containing "laminect", "colectom", "thoracotom", or "thoracosc" in the scheduled procedure name as free text, or the CPT codes. Patients were excluded if younger than 18 years, if the patient was not discharged at the time of data extraction, or if the surgery lasted <10 minutes. The latter was used as a safety net to exclude cancelled cases. If a single patient underwent multiple surgeries in the same dataset, only the first surgery was used. From the 533,408 procedures recorded in the PDW, 7,251 matched our criteria, resulting in datasets for laminectomy, colectomy, and thoracic surgery containing 2,328, 2,245, and 2,678 patients, respectively. This represents only 1.3% of the procedures recorded in the PDW because of the number dilution by all frequently performed minimally and non-invasive procedures, including endoscopies, cataract surgeries, interventional cardiology, and radiology, among

others. The features in the PDW are readily obtainable from the EHR and several come from tables specifically designed to assist with preoperative evaluation. For example, features for determining the possible presence of heart failure are drawn from multidimensional areas: medications, laboratory values, ICD codes, problem list, medical history, prior surgeries, echocardiogram results, and notes. To understand the data entry process and preoperative evaluations made to populate the PDW, essential to understand the generalizability of use in other institutions, we refer the reader to a recent publication summarizing this process[30].

Based on similar work on digital phenotype, this sample size and amount of data are sufficient to build reliable models. While some publications used significantly more patients (i.e. 16,552 unique patients to phenotype sepsis [23]; 134,252 for arthroplasties[27]), other authors published relevant results with smaller number of patients (608 patients for COVID-19 [19]; 300 patients for breast cancer [31], 105 patients for spine disease [32]) by compensating with increased number of data points.

Clinical endpoints

We believe that the best way to ensure the relevance of the phenotypes was to demonstrate their association with clinically important outcomes for the patient, the clinician, and the hospital. Even if a very solid segmentation had been achieved, based on high Silhouette and AMI scores, the real-world relevance of these phenotypes would remain low if they were not associated with clinically relevant outcomes, or if the association with these outcomes was lost when the phenotype was attributed to new patients.

To evaluate the clinical relevance of the phenotypes attributed to patients, we explored the association of each phenotype with five adverse outcomes: (1) in-hospital mortality; (2) 30-day mortality; (3)

reoperation within 30 days; (4) intensive care unit (ICU) admission; and (5) prolonged postoperative hospital length of stay (LOS). These adverse outcomes were not used to create phenotype, but only to validate their clinical relevance. 30-day mortality and in-hospital mortality were defined by a death recorded in the EHR within 30 days following surgery, and at hospital discharge, respectively [12]. Reoperation was defined as the occurrence of any surgery performed by the same service (i.e. general surgery for colectomy) within 30-days following surgery, to avoid considering unrelated surgery. A patient was defined as being admitted to ICU if any hour was spent in the ICU following surgery. Finally, prolonged hospital LOS was defined as postoperative LOS greater than the 90th percentile, established in the derivation dataset of each surgical cohort.

Patient features

Despite considering three distinct surgery-specific datasets, we first extracted the same dataset of 77 preoperative features (see supplementary material) for all three datasets. The features included demographics (e.g. weight), specific comorbidities (e.g., diabetes), labs, medication, and preoperative surgery or anesthesia features. These features were first selected by clinical experts' consensus (PLL, TW, MC) based on their availability in the preoperative setting and their potential clinical influence on postoperative evolution. Per our objectives, we aimed for preoperative risk stratification and excluded intraoperative features despite their established influence on postoperative outcomes [33].

Outcome-driven feature elimination

Unsupervised machine learning algorithms, such as k-means, typically attribute equal weight to all input features regardless of their clinical importance. For example, the feature "eye color", if included, could weigh as much as "diabetes" to segment into subpopulations. Therefore, we used an outcome-driven approach and only retain features presenting a statistically significant Pearson's correlation (P = 0.05) with

at least one of the clinical endpoints. The counterpart of this approach is the elimination of features correlated non-linearly to the endpoints. Categorical variables were one-hot encoded to calculate the correlation. We eliminated highly correlated features to avoid overweighting the condition linked to these features.

Data preprocessing

The data was preprocessed independently for each dataset by following the same approach. Values outside a physiological range (e.g., arterial pressure of 0- or 300-mm Hg) were considered registration artefacts and treated as missing values, similarly to recent publications[8]. If missingness was over 40%, a clinical consensus (PLL, TW) evaluated the relevance of keeping the feature (supplementary table 1). For example, left ventricular ejection fraction (LVEF) is rarely available prior to laminectomy but remains relevant. Multivariate imputation by chained equations was used to account for missing data [34]. Continuous features were normalized. The final surgery-specific datasets contained 34, 36, and 33 features for laminectomy, colectomy, and thoracic surgery, respectively (supplementary table 2 for included features, and supplementary table 3 for excluded features).

Separation into derivation and validation cohorts

Each surgery-specific dataset was separated into derivation and validation cohorts based on chronological occurrence to mimic the prospective attribution of phenotypes to the new patients to future patients (see supplementary figure 1). During anonymization of the data, the institution solely retains the year of occurrence. By losing this granularity, we could not precisely separate our cohorts with a fixed percentage and thus, we separated by using on the year to hold out between 30 and 40% of the patients in the validation cohort. The validation cohort remained untouched throughout model derivation. For each

validation cohort, we confirmed that the occurrence of all endpoints (adverse outcomes) was similar to the incidence found in the derivation datasets.

By adopting a chronological validation approach, we aimed at mimicking a prospective attribution of phenotypes to future patients, based on a model built on previous patients.

Derivation of phenotypes

Supplementary figure 1 provides an overview of the derivation and validation methods. In accordance with data preprocessing, we independently derived three distinct phenotyping models for each surgery, by using their respective datasets. We used a 10-iterations consensus clustering approach [35] with K-means[36]. This approach was previously used for clinical modelling[35], but we compared its performance to other segmentation strategies (DB-Scan, hierarchical descending, k-means) to obtain the optimal consistency and robustness (supplementary tables 4-5). The optimal number of clusters was established by using a combination of Silhouette score, normalized mutual info (NMI) score, homogeneity score, adjusted Rand index (ARI), consensus matrix heatmaps, pairwise-consensus values for all patients, and characteristics of the consensus cumulative distribution function plots (see supplementary figure 2). These results, combined with the clinical consensus, established that consensus k-means with three clusters was the optimal segmentation strategy.

Based on the results of our model exploration, a consensus k-means modelling with three clusters (k=3) was applied to all the patients from each derivation cohorts. A first exploration of the phenotype's characteristics and the association with adverse outcomes was explored.

Model validation

By definition, a consensus k-means algorithm will always succeed at segmenting a population in a pre-established number of clusters. To confirm that our model derived consistent phenotypes, we validated the model by exploring if prospectively attributed phenotypes maintained the similar phenotyping profile and their relation with adverse post-operative outcomes.

To account for the impracticality of clustering a new patient after consensus k-means, we used patients' features to establish which phenotype should be attributed to new patients. We applied a train-test split within the derivation and used phenotypes attribution to train a random forest to predict and attribute a phenotype to the patients within in the validation cohort [37] . Recent publications confirmed that the strategy of using patients' features to prospectively attribute digital phenotypes was effective, even if the algorithm explored were slightly different[37].

We compared the distributions of phenotypes across the derivation and validation sets. The inter-cluster difference for each endpoint was studied by clinical experts (PLL, TW, MC). The main characteristics of each phenotype were explored and compared to the main characteristics found within the derivation cohort. We explored the association between the phenotypes of the validation cohort, and the occurence of adverse outcomes. We used the area under the receiver operating characteristic curve (AUROC) to compare our approach to ASA score alone and we investigated a combination of our phenotypes with the ASA score (supplementary table 7). By considering this combination, we aimed to evaluate their complementary nature and explore potential synergies in improving patient risk stratification.

The ASA score was chosen as a suitable comparator in our study due to its widespread use and acceptance within the medical community. It served as an appropriate benchmark because, like our developed

phenotypes, each ASA score represents a distinct group of patients, sharing more than a common risk profile, but also health-related characteristics.

Results

The validation cohorts for laminectomy, colectomy, and thoracic surgery contained 999 patients (30.5% of total), 768 (34.2%), and 1,003 (37.5%), respectively. All results in this section refer to the validation cohort unless otherwise specified.

Patients

Table 1 presents patient characteristics in the validation sets for all three surgeries (see supplementary table 6 for derivation set characteristics). The median ASA score was 3, and scores of 2 and 3 were the most prevalent. Patients undergoing thoracic surgery had the most complications for all adverse outcomes recorded.

Derivation of phenotypes

Three phenotypes (k = 3) provided the optimal fit for all three surgical groups. After reordering phenotypes from low to high risk based on the relative occurrence of adverse outcomes, the low-risk phenotype *alpha* consistently grouped most patients (65% to 75% of the patients depending on the surgery). For each surgery, the distribution of each phenotype was compared between the derivation and the validation datasets to validate our prospective attribution (supplementary figure 2). Both colectomy and thoracic surgery presented a similar distribution, but a difference existed for laminectomy: the high-risk phenotype *gamma* was not attributed to any patients in the validation dataset despite representing 9% of the laminectomy derivation dataset. As further addressed in the discussion, the descriptive analysis revealed a significant difference in the comorbidity profile between the patients in the derivation and the validation datasets. These comorbidities were key factors to define the *gamma* phenotype.

Adverse outcomes

The occurrence of adverse outcomes increased consistently from the most common phenotype *alpha* to phenotype *beta* and *gamma*. The combined in-hospital mortality increased from 0.2% with phenotype *alpha* to 2.3% and 7.3% for phenotypes *beta* and *gamma*, respectively. Despite being attribute nearly five and seven times more often than the other phenotypes, the low-risk phenotype *alpha* presented a similar absolute count of adverse events (figure 1.a), resulting in a significantly lower rate of adverse events for each patient with phenotype *alpha* (figure 1.b). A similar progression existed across all surgeries and across all outcomes, with phenotype *alpha* systematically representing the largest cohort and at least two thirds of the patient. Only reoperation exhibited a less consistent progression (see supplementary figure 5).

Clinical characteristics of phenotypes

Given the notable differences between the three surgeries studied, we conducted individual exploration for each surgery to identify the most influential features contributing to the attribution of phenotypes. This approach facilitated a more nuanced understanding of the influential features within each surgical domain. Figure 2 summarizes (a) the two-by-two analysis of the phenotypes' characteristics for colectomy (laminectomy and thoracic surgery found in supplemental), and (b) the median and mean values for the most relevant features of each phenotype, and in each surgery. Figure 2 summarizes the following section, in which we present the key characteristics of each surgery-specific phenotypes.

Laminectomy

When compared to intermediate-risk phenotype *beta,* we notice that patients undergoing laminectomy with a digital phenotype *alpha* were older (69 vs 44 years old), scheduled for shorter surgery (185 vs 263 minutes), had not been hospitalized preoperatively (0 vs 2 days of preoperative hospitalization), and had

lower preoperative pain (2 vs 5 on a visual analog scale). Interestingly, the lower risk *alpha* phenotype had more respiratory flag (41% vs 20%) and CHF flag (13% vs 7%) than *beta* phenotype.

Colectomy

When compared to phenotype *beta,* phenotype *alpha* was attributed mostly to younger (58 vs 73 years old) and healthier patients (CHF flag: 11% vs 91%; respiratory flag: 26% vs 57%) undergoing an elective procedure. Both *alpha* and *beta* phenotypes had similar baseline preoperative heart rate (75 vs 76 bpm), had no total parenteral nutrition (0% for both) and no preoperative hospitalization. On the other hand, *gamma* phenotype comprised most non-elective surgeries and, when compared to the two other phenotypes, had more a longer pre-surgical hospitalization (median duration of 7 days), higher baseline heart rate (85 bpm) and increased use of parenteral nutrition (22%). They were usually younger and had less comorbidities than *beta* phenotype.

Thoracic surgery

Patients with phenotype *alpha* underwent elective surgery, were often the first case scheduled for the day (61%), had less cardiac or endocrine flags of comorbidities (12% and 8% respectively) and none had a pre-surgical stay. Phenotype *beta* was highly characterized by diabetes (100%) and cardiac flag (30%) and exhibited a mixture elective and non-elective cases. As seen with colectomy, phenotype *gamma* was mostly attributed to patients undergoing non-elective surgeries and had stayed at the hospital for a longer preoperative stay (median 10 days). They were more often female (65%), their heart rate was significantly faster (90 bpm) and they exhibited more pain in the preoperative setting (7 on VAS).

Comparison With ASA Score

Most patients with an ASA score of 2 were attributed phenotype *alpha* (82.2%) and, as ASA score increased, the proportion of higher-risk phenotype *gamma* also increased (ASA 4: 49%; ASA 5: 66.67%). Out of the 2 770 patients comprising the validation cohorts, 2 039 were attributed the low-risk phenotype *alpha*, of which 1 684 had the intermediate-risk ASA score of 3. On the other hand, most patients with an ASA score of 3 were classified as low-risk phenotype *alpha* (1267/1684; 75%). Figure 3 offers a visual representation of the relationship between phenotypes and ASA score.

Based on the AUROC, the association between all five adverse outcomes and the three-class phenotypes was either similar or slightly higher than the five-class ASA score (see figure 4). Phenotyping outperformed ASA score most significantly when predicting ICU admission (AUROC 0.76 vs. 0.71) and prolonged LOS (0.75 vs. 0.71). Reoperation was the most challenging outcome to predict for both approaches (AUROC 0.59 and 0.62).

Combination With ASA Score

Finally, the combination of ASA score and digital phenotyping outperformed either scores alone, when using AUROC as the comparing metric. The linear combination of phenotypes and ASA score created a total 15 distinct categories (ASA1 – phenotype *alpha*; ASA1 – phenotype *beta*; and so on; see supplementary table 7). Combining human insight and digital phenotyping improved prediction for all outcomes and all surgeries. This linear combination reached an AUROC of 0.91 for hospital mortality (phenotype 0.85; ASA: 0.84), and an AUROC of 0.80 for both ICU admission and prolonged hospitalization (phenotype 0.75 and 0.75; ASA: 0.71 and 0.71).

Discussion

In this study, we developed a method for preoperative risk stratification by using digital phenotypes, which created more homogeneous subgroups from a heterogeneous population. These subgroups shared common characteristics and an increased association with adverse outcome when compared with ASA score. This association was maintained when prospectively attributed to new patients.

We explored our method among three populations undergoing three frequently performed surgeries, specifically selected to cover a wide range of perioperative risks. Across all three surgical groups, the low-risk phenotype *alpha* was the most frequently attributed phenotype (over two thirds of the patient population), corresponding to what is statistically expected in a general population. The occurrence of adverse events increased progressively from the phenotype *alpha* to the higher-risk phenotype *gamma*. The consistency of these two finding across all three surgical populations supports the generalizability and scalability of the model to other surgical populations, which will have to be investigated in future publications.

This methodology is in contrast with most published strategies, in which supervised machine learning models are trained to predict specific complications. A similar strategy of using supervised models predicting mortality and hospital LOS could have been explored, before categorizing the predictions in three groups: low, moderate, and elevated risk [38]. Nonetheless, digital phenotyping is an alternate, complementary strategy that merits exploring due to its ability to track digital signatures and inform the clinician of the key characteristics of each phenotype instead of solely providing a risk score, built by forcing the model into selecting the features relevant to pre-defined outcomes and thus lose the ability to predict other distinct relevant outcomes unless the model is retrained. By phenotyping patients, we can theoretically predict unrelated complications, for example post-operative anemia and patient

readmission, which are two additional outcomes to explore in future works. Moreover, phenotyping can be used to personalize new therapies, as seen in sepsis and psychiatric conditions, or help match new patients with previous similar patients who could offer mentoring and support. This has been successfully explored in other fields of medicine characterized by heterogenous populations, including sepsis, in which digital phenotype based on labs and organ dysfunction predicted mortality and relevance of fluid therapies[23]. Similarly in psychiatric conditions, digital phenotypes built from symptoms and ambulation was used to predict relapse and dangerousness, guiding requirement for hospitalization [39-41]. Overall, we suggest that digital signatures and phenotypes have the potential to yield a higher-level of understanding of our patients, instead of solely focusing on classic outcomes like mortality and LOS.

ASA score, one of the most widely used score, could be considered a phenotype considering that is is based on patient's characteristics and is not directly of percentage of risk of mortality. Kowing that a patient has an ASA score of 1 is an efficient way to inform an anesthesiologist on the patient's status, but ASA score of 3, more commonly attributed, is much more heterogeneous. This challenge is partly solved with digital phenotypes, with which the low-risk phenotype *alpha* is the most frequentlyattributed cluster. Based on the ROC curve profiles (figure 4), patients with ASA 2 and 3 scores benefit the most from digital phenotyping due to their heterogeneity and common use. Compared to ASA score, digital phenotyping eliminates the reliance on experts and should reduce inter-user variability by being automatically attributed from readily extractable data. Phenotyping can be used by itself or in combination with human evaluation (i.e. with ASA scoring). In clinical setting, this could be translated in an automated attribution of a phenotype based on EHR data, providing a first automated insight on the patient trajectory, and to update the patient's risk stratification when the anesthesiologist evaluates the patient and assigns an ASA score. More specifically, a primary evaluation could be made before the decision of surgery is made, potentially influencing the surgeon's and the patient's choice to proceed. This primary evaluation would

be updated after preoperative evaluation, and after any updates until the surgery. Eventually, a model integrating intraoperative events would further improve the model.

Interestingly, the most predictive features varied according to the surgery, an intuitive clinical concept learnt automatically by our models. It is important to mention that, after thorough discussion among the authors, we decided to exclude race and ethnicity from our models. Because race and ethnicity may be colinearly associated with many other features, significance and implications can be misattributed. Specifically, including race and ethnicity in the model could perpetuate racial behaviours or inequities, as the model is only able to learn from current observed data. This topic arose recently in a study that used a machine learning algorithm to predict transfusion risk [42-44].

When prospectively attributing phenotypes in the laminectomy validation dataset, no patients were attributed a phenotype *gamma*, despite being attributed to 9.2% of the patients in the derivation dataset (supplementary figure 2). A descriptive analysis of these patients revealed that on average a patient in the derivation cohort with a phenotype *gamma* had 3.5 comorbidities while no patients presented any comorbidities in the laminectomy validation cohort. The profile of the population significantly changed during the COVID-19 pandemic for this kind of elective functional surgery and, by using a temporal split to create our validation cohort, our results showed accordingly that highly comorbid patients were not operated on during the pandemic. Other approaches could have been adopted to compensate for the changes of population during COVID-19. However, we considered it of higher interest and more generalizable to create the phenotypes within a “usual” population without pandemic-related changes and to analyse how these phenotypes subsequently performed in a population modified by COVID-19.

Throughout this research, the methodology has been selected to be applicable to any surgical population. With minor modifications, the same method can be translated to 1) identify and attribute phenotypes prior to any surgery frequently performed, 2) include new features, and 3) predict other relevant outcomes. Adopting a temporal validation supports the applicability of the algorithm over time and as observed with laminectomy, react adequately despite significant changes in the surgical population. However, both the derivation and validation cohorts originated from the same database, and generalizability to other populations needs to be confirmed. From a machine learning standpoint, the low occurrence and limited number of adverse outcomes remain a challenge both during derivation and validation of phenotyping approaches. All the analysis and conclusions rely on the veracity of the data extracted; as such, our results are vulnerable to any possible artefacts and errors in our database. As an example of such an artefact, a patient who died in a hospital outside of the UCLA healthcare system would not necessarily be recorded as deceased. In our current model, the variation in the complexity of the procedure was not grasped despite a significant influence on postoperative outcome. Apart from a longer planned surgery duration, no features could help distinguish between a wedge and pneumonectomy, or a single-level laminectomy and a multiple-level surgery with neuromuscular comorbidities. Finally, as for any machine learning algorithms developed, the clinical application of phenotyping requires an EHR.

Conclusion

In this analysis of patients undergoing laminectomy, colectomy, and thoracic surgery, we confirmed the relevance of digital phenotyping as a tool for risk stratification. For each surgery group, we obtained three distinct preoperative digital phenotypes with distinct clinical characteristics and postoperative care trajectories. Future investigations will apply this method to other surgical groups, and validate in another institution, and could include features collected intraoperatively to influence the digital phenotype during surgery.

References

1. Nepogodiev D, Martin J, Biccard B, Makupe A, Bhangu A, National Institute for Health Research Global Health Research Unit on Global S. Global burden of postoperative death. *Lancet*. Feb 2 2019;393(10170):401. doi:10.1016/S0140-6736(18)33139-8
2. Weiser TG, Regenbogen SE, Thompson KD, et al. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet*. Jul 12 2008;372(9633):139-144. doi:10.1016/S0140-6736(08)60878-8
3. Pearse RM, Harrison DA, James P, et al. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Crit Care*. 2006;10(3):R81. doi:10.1186/cc4928
4. Le Manach Y, Collins G, Rodseth R, et al. Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and Validation. *Anesthesiology*. Mar 2016;124(3):570-9. doi:10.1097/ALN.0000000000000972
5. Stavem K, Hoel H, Skjaker SA, Haagensen R. Charlson comorbidity index derived from chart review or administrative data: agreement and prediction of mortality in intensive care patients. *Clin Epidemiol*. 2017;9:311-320. doi:10.2147/CLEP.S133624
6. Hofer IS, Lee C, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model to predict postoperative mortality, acute kidney injury, and reintubation using a single feature set. *NPJ Digit Med*. 2020;3:58. doi:10.1038/s41746-020-0248-0
7. Hackett NJ, De Oliveira GS, Jain UK, Kim JY. ASA class is a reliable independent predictor of medical complications and mortality following surgery. *Int J Surg*. Jun 2015;18:184-90. doi:10.1016/j.ijsu.2015.04.079
8. Horvath B, Kloesel B, Todd MM, Cole DJ, Prielipp RC. The Evolution, Current Value, and Future of the American Society of Anesthesiologists Physical Status Classification System. *Anesthesiology*. Nov 1 2021;135(5):904-919. doi:10.1097/ALN.0000000000003947
9. Kwa CXW, Cui J, Lim DYZ, Sim YE, Ke Y, Abdullah HR. Discordant American Society of Anesthesiologists Physical Status Classification between anesthesiologists and surgeons and its correlation with adverse patient outcomes. *Sci Rep*. May 2 2022;12(1):7110. doi:10.1038/s41598-022-10736-5
10. Tollinche LE, Yang G, Tan KS, Borchardt R. Interrater variability in ASA physical status assignment: an analysis in the pediatric cancer setting. *J Anesth*. Apr 2018;32(2):211-218. doi:10.1007/s00540-018-2463-2
11. Hill BL, Brown R, Gabel E, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *Br J Anaesth*. Dec 2019;123(6):877-886. doi:10.1016/j.bja.2019.07.030
12. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality. *Anesthesiology*. Oct 2018;129(4):649-662. doi:10.1097/ALN.0000000000002186
13. Lee CK, Samad M, Hofer I, Cannesson M, Baldi P. Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality. *NPJ Digit Med*. Jan 8 2021;4(1):8. doi:10.1038/s41746-020-00377-1
14. Misic VV, Gabel E, Hofer I, Rajaram K, Mahajan A. Machine Learning Prediction of Postoperative Emergency Department Hospital Readmission. *Anesthesiology*. May 2020;132(5):968-980. doi:10.1097/ALN.0000000000003140

15. Tseng PY, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care*. Jul 31 2020;24(1):478. doi:10.1186/s13054-020-03179-9
16. Gray GM, Ahumada LM, Rehman MA, et al. A machine-learning approach for decision support and risk stratification of pediatric perioperative patients based on the APRICOT dataset. *Paediatr Anaesth*. May 21 2023;doi:10.1111/pan.14694
17. Wongtangman K, Aasman B, Garg S, et al. Development and validation of a machine learning ASA-score to identify candidates for comprehensive preoperative screening and risk stratification. *J Clin Anesth*. Aug 2023;87:111103. doi:10.1016/j.jclinane.2023.111103
18. Zhang L, Fabbri D, Lasko TA, Ehrenfeld JM, Wanderer JP. A System for Automated Determination of Perioperative Patient Acuity. *J Med Syst*. May 30 2018;42(7):123. doi:10.1007/s10916-018-0977-7
19. Data Science Collaborative G. Differences in clinical deterioration among three sub-phenotypes of COVID-19 patients at the time of first positive test: results from a clustering analysis. *Intensive Care Med*. Jan 2021;47(1):113-115. doi:10.1007/s00134-020-06236-7
20. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. *Nat Biotechnol*. May 2015;33(5):462-3. doi:10.1038/nbt.3223
21. Oellrich A, Collier N, Groza T, et al. The digital revolution in phenotyping. *Brief Bioinform*. Sep 2016;17(5):819-30. doi:10.1093/bib/bbv083
22. Ferreira JC, Patino CM. Types of outcomes in clinical research. *J Bras Pneumol*. Jan-Feb 2017;43(1):5. doi:10.1590/S1806-37562017000000021
23. Seymour CW, Kennedy JN, Wang S, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *JAMA*. May 28 2019;321(20):2003-2017. doi:10.1001/jama.2019.5791
24. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital Phenotyping of Mental Health using multimodal sensing of multiple situations of interest: A Systematic Literature Review. *J Biomed Inform*. Feb 2023;138:104278. doi:10.1016/j.jbi.2022.104278
25. Jayakumar P, Lin E, Galea V, et al. Digital Phenotyping and Patient-Generated Health Data for Outcome Measurement in Surgical Care: A Scoping Review. *J Pers Med*. Dec 15 2020;10(4)doi:10.3390/jpm10040282
26. Grant RW, McCloskey J, Hatfield M, et al. Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles. *JAMA Netw Open*. Dec 1 2020;3(12):e2029068. doi:10.1001/jamanetworkopen.2020.29068
27. Ranti D, Warburton AJ, Hanss K, Katz D, Poeran J, Moucha C. K-Means Clustering to Elucidate Vulnerable Subpopulations Among Medicare Patients Undergoing Total Joint Arthroplasty. *J Arthroplasty*. Dec 2020;35(12):3488-3497. doi:10.1016/j.arth.2020.06.063
28. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res*. Dec 16 2016;18(12):e323. doi:10.2196/jmir.5870
29. Hofer IS, Gabel E, Pfeffer M, Mahbouba M, Mahajan A. A Systematic Approach to Creation of a Perioperative Data Warehouse. *Anesth Analg*. Jun 2016;122(6):1880-4. doi:10.1213/ANE.0000000000001201

30. Kamdar NV, Huverserian A, Jalilian L, et al. Development, Implementation, and Evaluation of a Telemedicine Preoperative Evaluation Initiative at a Major Academic Medical Center. *Anesth Analg*. Dec 2020;131(6):1647-1656. doi:10.1213/ANE.0000000000005208
31. Delrieu L, Hamy AS, Coussy F, et al. Digital phenotyping in young breast cancer patients treated with neoadjuvant chemotherapy (the NeoFit Trial): protocol for a national, multicenter single-arm trial. *BMC Cancer*. May 4 2022;22(1):493. doi:10.1186/s12885-022-09608-y
32. Cote DJ, Barnett I, Onnela JP, Smith TR. Digital Phenotyping in Patients with Spine Disease: A Novel Approach to Quantifying Mobility and Quality of Life. *World Neurosurg*. Jun 2019;126:e241-e249. doi:10.1016/j.wneu.2019.01.297
33. Lamarche Y, Elmi-Sarabi M, Ding L, Abel JG, Sirounis D, Denault AY. A score to estimate 30-day mortality after intensive care admission after cardiac surgery. *J Thorac Cardiovasc Surg*. May 2017;153(5):1118-1125 e4. doi:10.1016/j.jtcvs.2016.11.039
34. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16(3):219-242. doi:10.1177/0962280206074463
35. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. Jun 15 2010;26(12):1572-1573. doi:10.1093/bioinformatics/btq170
36. MacQueen JB. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967;1:281-297.
37. Alyousef AA, Nihtyanova S, Denton C, Bosoni P, Bellazzi R, Tucker A. Nearest Consensus Clustering Classification to Identify Subclasses and Predict Disease. *J Healthc Inform Res*. 2018;2(4):402-422. doi:10.1007/s41666-018-0029-6
38. Jalali A, Lonsdale H, Do N, et al. Deep Learning for Improved Risk Prediction in Surgical Outcomes. *Sci Rep*. Jun 9 2020;10(1):9289. doi:10.1038/s41598-020-62971-3
39. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology*. Jul 2018;43(8):1660-1666. doi:10.1038/s41386-018-0030-z
40. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic Review of Digital Phenotyping and Machine Learning in Psychosis Spectrum Illnesses. *Harv Rev Psychiatry*. Sep/Oct 2020;28(5):296-304. doi:10.1097/HRP.0000000000000268
41. Kleiman EM, Turner BJ, Fedor S, et al. Digital phenotyping of suicidal thoughts. *Depress Anxiety*. Jul 2018;35(7):601-608. doi:10.1002/da.22730
42. Brittany N. Burton CC, Jennifer Lucero, Maxime Cannesson. Personalized Surgical Transfusion Risk Prediction: Comment. *Anesthesiology*. 2022;XXX::00-00.
43. Lou SS, Liu H, Lu C, Wildes TS, Hall BL, Kannampallil T. Personalized Surgical Transfusion Risk Prediction Using Machine Learning to Guide Preoperative Type and Screen Orders. *Anesthesiology*. Jul 1 2022;137(1):55-66. doi:10.1097/ALN.0000000000004139
44. Sunny S. Lou TSW, Bruce L. Hall, Michael S. Avidan, Thomas Kannampallil. Personalized Surgical Transfusion Risk Prediction: Reply. *Anesthesiology*. 2022;XXX:00-00.

Tables

Table 1. Patients' characteristics and occurrence of adverse outcomes in the validation sets for all investigated surgeries.

	Laminectomy (n=999)	**Colectomy (n = 768)**	**Thoracic (n=1003)**
Preoperative characteristics			
Age Mean (SD)	62 (16)	58 (16)	60 (17)
Female N (%)	589 (59.0%)	400 (52.1%)	525 (52.3%)
Race and ethnicity[1]			
African American	54 (5%)	48 (6%)	57 (6%)
Caucasian	636 (64%)	436 (57%)	591(59%)
Asian American	101 (10%)	86 (11%)	125(12%)
Other	132 (13%)	129 (17%)	147 (15%)
Unknown	76 (8%)	69 (9%)	83 (8%)
Hispanic	113 (11%)	130 (17%)	136 (14%)
Non-Hispanic	892 (80%)	587 (76%)	791 (79%)
Unknown	84 (9%)	51 (7%)	76 (8%)
ASA score Median [IQR]	3 [2-3]	3 [2-3]	3 [3-3]
Weight Mean (SD)	82 (19)	76 (19)	75 (19)
Body mass index Mean (SD)	28 (6)	26 (6)	26 (5)
Max preoperative pain Median [IQR]	0 [0-6]	0 [0-0]	0 [0-2]
Diabetes N (%)	344 (26.5%)	167 (23.3%)	195 (21.4%)
Smoking N (%)	60 (6.3%)	47 (6.3%)	49 (5.1%)
Respiratory pathology N (%)	334 (36.2%)	202 (28.2%)	348 (38.2%)
Chronic heart failure N (%)	109 (11.8%)	81 (11.3%)	147 (16.1%)
Ischemic heart disease N (%)	158 (17.1%)	113 (15.8%)	181 (19.8%)
Last ICU stay (hours) Median [IQR]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Preoperative LOS (days) Median [IQR]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Total parenteral nutrition in previous 48h - N (%)	1 (0.1%)	27 (3.5%)	5 (0.5%)
Elective surgery N (%)	835 (83.6%)	644 (83.9%)	803 (80.1%)
Preoperative heart rate Mean (SD)	74 (13)	77 (15)	75 (16)
Preoperative mean arterial pressure - Mean (SD)	96 (12)	91 (12)	90 (12)
Scheduled minutes Median [IQR]	180 [165-240]	240 [180-300]	240 [180-300]

Postoperative adverse outcomes			
30-day mortality N (%)	8 (0.8%)	4 (0.5%)	21 (2.1%)
Inpatient mortality N (%)	9 (0.9%)	7 (0.9%)	20 (2.0%)
30-day reoperation N (%)	38 (3.8%)	18 (2.3%)	53 (5.3%)
ICU admission N (%)	149 (14.9%)	44 (5.7%)	271 (27.0%)
Prolonged hospital LOS[2] N (%)	98 (9.8%)	51 (6.6%)	94 (9.4%)

For further details regarding the mapping of these characteristics and the electronic health record, please consult supplementary material; [1] Race and ethnicity is presented in this table but was not used as a feature for the derivation and validation of the model. The underlying reasons are described in the discussion; [2]Defined by hospital length of stay higher than the 90th percentile, as measured in the derivation cohorts. Abbreviations: ASA: American Society of Anesthesiology; BMI: body mass index; LOS: length of stay.

Table 2. Incidence of adverse outcomes for each phenotype in surgery-specific validation datasets

	30-day mortality	Inpatient mortality	30-day reoperation	ICU admission	Prolonged hospital LOS[1]
Laminectomy					
Alpha (*N* = 743)	0.1%	0.1%	2.4%	4.4%	3.1%
Beta (*N* = 256)	2.7%	3.1%	7.8%	45.3%	29.3%
Gamma (*N* = 0)	-	-	-	-	-
Colectomy					
Alpha (*N* = 607)	0.0%	0.0%	1.6%	2.1%	3.0%
Beta (*N* = 44)	2.3%	2.3%	0.0%	4.5%	6.8%
Gamma (*N* = 117)	2.6%	5.1%	6.8%	24.8%	25.6%
Thoracotomy					
Alpha (*N* = 689)	0.7%	0.4%	4.4%	17.1%	4.2%
Beta (*N* = 129)	0.8%	0.8%	2.3%	28.7%	13.2%
Gamma (*N* = 185)	8.1%	8.6%	10.8%	62.7%	25.9%
Global					
Alpha (*N* = 2039)	0.3%	0.2%	2.8%	8.0%	3.4%
Beta (*N* = 429)	2.1%	2.3%	5.4%	36.1%	22.1%
Gamma (*N* = 302)	6.0%	7.3%	9.3%	48.0%	25.8%

[1]Defined by hospital length of stay higher than the 90e percentile, as measured in the derivation dataset

Figures Caption

Figure 1. Chord diagrams for the (a) absolute and (b) relative occurrence of adverse care trajectories for phenotype *alpha* (low-risk), phenotype *beta* (intermediate risk), and phenotype *gamma* (high-risk), for all surgery-specific phenotypes combined.

Figure 2. Description of the most important features for each surgery, based on (a) random forest algorithm for prospective analysis and (b) heatmap grouping the surgery-specific discriminative features among the phenotypes.

Figures 1. b) present the median value (if continuous) and the percentage of occurrence (if binary) for each of the most discriminative features. The features were considered highly discriminative if the normalized standard deviation across the three phenotypes was elevated. The color coding is based on the quintile of the value. CHF: congestive heart failure; ICU: intensive care unit; LOS: length of stay; VAS: visual analog scale

Figure 3. Chord diagram comparing the attribution of ASA score and phenotypes.

This figure depicts that low-risk phenotype alpha was the most frequently attributed phenotypes and was constituted mostly by patients with ASA score of 2 and 3. High-risk phenotype gamma was mostly constituted by patients with ASA score of 3 and 4.

Figure 4. ROC curves for the prediction of clinical outcomes, based on digital phenotype, ASA score and the linear combination of both. a) hospital mortality; b) 30-day mortality; c) reoperation at 30 days; d) ICU admission; e) hospital length of stay over 90e percentile.

Bootstrap analysis with 100 iteration was used to determine the 95% confidence interval, shown on the figures as the shade around each line. The detailed approach for combining digital phenotype with ASA score is described in supplementary table 7. ASA: American Society of Anesthesiologist 'score; AUC: area under the curve; ICU: intensive care unit; LOS: length of stay.