What do we know about contributing factors for “never events” in operating rooms? A machine learning analysis

Dana Arad, MSN 1, 3, Ariel Rosenfeld, PhD 2, Racheli Magnezi, PhD 1

1 Department of Management, Health Management Program, Faculty of Sciences, Bar-Ilan University, Israel

2 Department of Information, Faculty of Science Bar-Ilan University, Israel

3 Patient Safety Division, Ministry of Health, Israel

Corresponding author:

Dana Arad, RN, BA, MSN, ACNP

[danaarad@gmail.com](mailto:danaarad@gmail.com)

**Acknowledgment:**

Acknowledgement: This study was funded by the Medical Research Fund for health services – Jerusalem, grant #MOHIG 14-2019

**What do we know about contributing factors for “never events” in operating rooms? A machine learning analysis**

Keywords: “Never Event,” surgery department, machine learning, contributing factors

**Abstract**

**Background**

A Surgical “Never Event” (NE) is a preventable error. Various factors contribute to the occurrence of wrong site surgery and retained foreign item, but little is known about their quantified risk in relation to surgery's characteristics. Our study uses machine learning to reveal factors and quantify their risk.

**Methods**

We used data from 9234 observations on safety standards and 101 Root-Cause Analysis from actual NEs, and utilized three Random Forest supervised machine learning models. Using a standard 10-cross validation technique, we evaluated the model's metrics, and through Gini impurity, we measured the impact of factors thereof to occurrence of the two types of NEs.

**Results**

We identified 24 contributing factors in six surgical departments. Two had an impact of >900% in Urology, Orthopedics and General Surgery, six had an impact of 0–900% in Gynecology, Urology and Cardiology, and 17 had an impact of <0%.

Factors' combination revealed 15-20 pairs with an increased probability in five departments: Gynecology:875–1900%; Urology: 1,900:2,600%; Cardiology:833–1,500%; Orthopedics:1,825–4,225%; and General Surgery:2,720–13,600%.

Five factors affected the occurrence of wrong site surgery (-60.96–503.92%) and five of retained foreign body (-74.65–151.43%), three of them overlapping: two nurses (66.26–87.92%), Surgery length<1 hour (85.56–122.91%), Surgery length 1-2 hours (-60.96–85.56%).

**Conclusions**

The use of machine learning has enabled us to quantify the potential impact of risk factors for wrong site surgeries and retained foreign items, in relation to surgery's characteristics, which in turn suggests tailoring the safety standards accordingly.

**Trial registration number:** MOH 032-2019

**BACKGROUND**

Adverse medical events can lead to significant morbidity and mortality and increase healthcare expenditures. [1] A Never Event (NE) is an unacceptable adverse event, both preventable and unjustified, and should be reduced to zero through quality improvement. [2 -3]. Major NEs in perioperative care include incorrect surgery sites and foreign items retained in patients following surgery. [4-5].

The human factors approach recognizes that human error is often the result of a combination of both individual surgeon factors and work system factors, [6] which makes human error the main contributing factor to NEs. [7-8]. Human error includes surgeon distraction, [9] lack of situational awareness of the surgical team to possible error, and miscommunication among team members. [10] Additionally, institutional factors, such as increased workload and clinician pressure, create a work climate that is not conducive to meeting the standards required to maintain patient safety [11] and effective teamwork. [12]

Currently, there are two essential international standards aiming to reduce NE occurrence: 1) the WHO Surgical Safety Checklist; [13] and 2) surgical counts of all items used during the surgery. [14]

Yet, partial compliance, unstandardized implementation of these standards, [15] and other possible unknown factors keep the incidence of NEs unchanged. [16] In Israel, the incidence of retained foreign items during surgery is 3.2 in every 100,000 surgeries. [17] The incidence for wrong site procedure is unclear, but is generally estimated as 1 in every 100,000 surgeries.

This study adopts a machine learning (ML) approach [18] to identify currently unknown contributors to NE occurrence. Previous studies leveraging ML methods in healthcare have demonstrated the benefits of analyzing and revealing non-trivial insights from diverse data types when compared to traditional methods. [19] To the best of our knowledge, this is the first study to use ML methods to identify potential contributing factors to the occurrence of NEs in ORs.

**METHODS**

**Study Design**

We utilized a supervised ML method called Random Forest (RF), [20] incorporating the popular Extra Tree classifier. [21]

RF is an ensemble learning method that trains multiple “simple” decision tree models and merges them to achieve a more accurate and stable prediction. The use of RF entails several desired elements needed for properly conducting the analysis for this study. First, RFs are used to rank the importance of features in a natural way. Specifically, the importance of features can be determined by examining to what extent the tree nodes using a feature reduce the impurity (i.e., the uncertainty in classification) across all “trees in the forest. Second, RFs are known to cope well with imbalanced datasets (as is the case in this study), and avoid overfitting the data. Finally, RFs compared favorably with several other supervised ML algorithms we tested using our data, including popular deep neural networks and support vector machines (SVMs). It is worthwhile mentioning that RFs have been used extensively in the medical field for clinical risk prediction, [22-23] among other applications.

Safety Standards used in the OR (surgical safety checklists and surgical counts) were divided into safety verification at three distinct time periods, pre-procedure, sign in and time out, [24]and addressed incorrect surgery site errors, which we will define as type A errors. Surgical counts were divided into three separate counts throughout the surgery to address retained foreign body errors, which we will define as type B errors: prior to skin incision; initiation of closure of fascia/cavity; and following skin closure. [25] In addition, we added general features such as the name of the hospital, length of surgery, patient’s gender and age, surgeon’s specialty, number of physicians and nurses present during surgery.

**Data Collection and Annotation**

Data were collected from 29 Israeli hospitals and consisted of two types of data entries: *observations* of 9,234 surgeries performed between January 2018 and February 2019 in which no NE occurred in the surgeries observed, and *root cause analyses* (RCA) of 101 NEs that occurred between January 2016 and February 2020 in the examined hospitals.

Observations

Initiated by the supervisory arm of the Israeli Ministry of Health, passive observations by medical students, physicians, nursing students or RNs are routinely performed in ORs. Observers for this study underwent an eight-hour long training that included simulations. In each OR, at least two observers passively observed randomly selected surgeries, and recorded and annotated the surgery process using a pre-defined set of features. Observations were then transferred to a central database and routinely assessed for variability and reliability. Overall, 9,234 observations were conducted. Each observation was translated into a 93-feature long vector, representing characteristics of the surgery (Appendix 1). To maintain reliability, entries with greater than 5% discordance among annotators in one OR were discarded (<1%).

Root Cause Analyses (RCA)

RCAs were performed in response to NEs that occurred between January, 2016 and February, 2020. Overall, we reported 101 NEs: 49 of Type A and 52 of Type B. The obtained RCAs were manually annotated by the authors using the same 93-feature-long representation used to characterize the observations. However, unlike the observations, RCAs were performed *retrospectively* and, thus, a significant portion of the features was missing and could not be obtained. Specifically, up to 40% of all other feature values were missing, a challenge we address further on.

**Pre-Processing and Analysis Technique**

As some features were non-binary (e.g., patient age, length of surgery), we first discretized them, resulting in 250 binary features. This and subsequent steps were performed using a designated Python 3 program implemented by the authors that uses the standard scikit-learn ML package (https://scikit-learn.org/stable).

Examination of the 40% missing feature values revealed that most were strongly dependent on the NE type. Namely, for type A NEs, features that were assumed to be more related to NEs of type B were not investigated and vice versa. For example, for an NE on which the wrong hand was operated, there was no indication as to whether the surgeon scanned the surgical cavity for retained surgical items before closure. To mitigate this artifact, we used the popular iterative data imputation approach [26] where we predicted the value of each missing value while relying on the present features and available examples. Specifically, using the entire dataset, each missing value was estimated using a standard Decision-Tree Regressor.

In addition, balancing steps were taken to cope with the high imbalance of the dataset. Specifically, with over 9,000 observations and only 101 NEs, we adopted a cost-sensitive training approach [27] whereby our model adjusted for prediction mistakes on the minority class (NEs) by an amount proportional to how underrepresented it was (here, approximately 90 times under-represented).

We implemented three RF models using our data: *Model 1* for differing between observations and NEs; *Model 2* for differing between observations and NEs Type A; and *Model 3* for differing between observations and NE Type B. We used a standard 10-cross validation technique to evaluate the model’s metrics and adopted the standard Gini impurity [28] measure to estimate the importance of features and their combination in our models. Intuitively, Gini impurity captures the “noise” in a set by measuring of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the set. Feature importance ranking was done using the trained RF models and we reported the change in NE occurrence probability given the entire data set. We considered each feature separately and calculated the probability of NE occurrence when that feature assumed the value True as compared to the value False.

The study was approved by the University's and Ministry of Health Ethics Committee (MOH 032-2019).

**RESULTS**

The majority of NEs (62.32%) occurred in six main departments: General Surgery, 19 (18.81%); Gynecology, 17 (16.83%); Orthopedics, 16 (15.84%); Cardiac and Cardiothoracic 15 (14.85%); Ophthalmology 8 (7.92%); and Urology, 7 (6.93%) (Table 1). Therefore, our analysis focused on the occurrence of NEs in these six departments.

Table 1: Characteristics of the data set according to surgical specialty

Table 2: Characteristics of patients and surgery in the dataset

Our three RF models demonstrated good performance, exhibiting an Area Under the Curve (AUC) between 0.81 and 0.85. AUC is interpreted as the probability that our model will rank a randomly chosen positive instance higher than a randomly chosen negative one. As such, our models can be considered relatively strong.

**Feature Importance**

Starting with Model 1, Table 3 presents the top contributing features to the occurrence of NEs (of both types combined) in the six departments along with the associated probability change.

Table 3: Top 14 contributing features for the six examined departments

The top 14 contributing features varied significantly across departments, and there was no single feature set which was consistently more informative across all operations in predicting NEs. For example, feature [C], **Discrepancy in second count,** varied significantly across departments (160% to 1,950%). Feature [B], **Surgery is paused because of discrepancy in third count,** appeared in four of the six departments, and the associated probability change varied dramatically as well, between 269% and 1,540%. There were 10 features that consistently decreased the chance of an NE, including [F]; **Surgeon scans the cavity/fascia before closure during the second count**, which affected five out of six departments, was consistent in its probability change between 65%–100%. Features [I], [J], [ K], [L], [M], and [N] decreased the chances of NEs between 2%–100% in three departments. Three features, [A], **Discrepancy in absorbing materials**, [E], **Surgery time > 4 hours,** and [G], **Surgery time < 1 hour** appeared just once across departments with a medium impact on NE occurrence.

Analysis of the results per department shows a variation among contributing features. For example, in Ophthalmology, the probability was consistently -100% in five features, while in General Surgery, two features that increased the probability of an error varied between 1,168–1,283%: features [B] **Surgery is paused because of discrepancy in third count;** and **[C] Discrepancy in second count.** In Orthopedics, those same two features, [**B**] and [**C**], increased the probability of error (1,540–1,950%). Three features decreased the probability of error: [F] **Surgeon scans the cavity/fascia before closure;** [H] **Second count is performed before closure of fascia/cavity; and** (I) **Procedure type is compared to the one written in patient's file,** -65 to -87%.

**Effects of Feature Combinations**

In the following analysis (Figure 1), we examine the effects of paired features, i.e., features that occur together in the data.

Figure 1: Effect of two features’ combination on prediction by surgical department

Interestingly, in General Surgery, there were 14 feature combinations that caused a probability change of 13,600% (Figure 1A). In comparison, the single feature analysis (Table 3) revealed a probability change of 1,287% and 1,168%, surprisingly by two features that were not part of the 14 feature combinations identified here.

In Figure 1A (Gynecology), the effect of every feature combination is associated with a probability change of 1,000–2,000%. In the single feature analysis (Table 2), the effect of two of the features separately was <900%, and the rest lagged behind with <150%. In Urology (Figure 1B), results show there were dozens of pairs with an effect of 1,900–2,500%, while the effect of a single feature had <1150% effect on error. In General Surgery (Figure 1E), the accumulated effect of two features together showed a dozen pairs with an effect of 1,900–4,200%, while the effect of a single feature had an <1,950% indication on error, and the rest even lower percentages.

**Features Affecting Types A and B**

Turning to Models 2 and 3, there is an overlap in three of the top five contributing features to Types A and B errors: 1) the presence of two nurses during the surgery predicts a greater occurrence of Type A (66%) and Type B (85%); 2) an operation < 1 hour had a greater occurrence of Type A (122%), and Type B (87%); and 3) when the operation lasted between one to two hours, both Types A and B were less frequent, decreasing by 60% and 74%, respectively. The surgical department that was most affected regarding the occurrence of Type A NEs was Ophthalmology, with a prevalence of 504%, while General Surgery was associated with a decrease of 63% in Type A (Figure 2). For Type B, the two remaining features were staff driven; the feature “more than three surgeons” was associated with an increased prevalence of Type B (122%), while “two surgeons” was associated with a decreased prevalence of 52% with Type B (Figure 3).

Figure 2: Features affecting the wrong site surgery (Type A)

Figure 3: Features affecting retained foreign item during surgery (Type B)

**DISCUSSION**

Surgical errors are a serious public health problem and uncovering their causes is challenging. [29] In this study, we aimed to uncover contributing factors to NEs by using ML methods to identify heretofore unknown contributors, since ML automatically looks for patterns not seen by classic methods. [30–32]

Despite the widespread use of the surgical safety checklist and strict surgical counts, the prevalence of NEs has not decreased significantly since their widespread implementation. [33–34] The human factor, and not system error, has been identified as the main contributing factor to NEs. [34–35] For example, in one study using an analysis and classification system, 628 human factors were divided into four categories that influenced NEs: preconditions for action, unsafe actions, oversight and supervisory factors, and organization influences. [24] Additional studies have identified lack of communication and lack of empirical evidence as barriers to the implementation of the universal safety standards. [28-29] Some studies have suggested that counting alone is insufficient and even when declared correct, there have been items left in the patient, [38–39] mostly in the abdomen and pelvis. [14,31] This may explain our higher probability of Type B error in General Surgery and Urology, which involve those regions.

We further analyzed paired contributing factors representing the relative risk in the OR’s complex work environment, when the graded risk increased compared to single feature analysis. For example, in Orthopedics, discrepancy in the count in combination with a surgery length of 1–2 hours increased the chances for an NE, what can be explained by partial compliance with the standards. In shorter surgeries, the staff rushes and skips some phases of the checklists [41] and the complex sets used challenges the counts. [34, 42]

We found that the occurrence of wrong site surgery increases in Ophthalmology during short surgeries and when two nurses are present. Its occurrence decreased in general surgery. This increased risk in could be due to the difficulty of performing a time out because the surgeons have antiseptic hands and cannot review charts, or perhaps doing so is not made a priority. [33] The decrease in general surgery could be explained by better implementation of the time out process in that specialty. [44–45]

One of the main factors contributing to the occurrence of NEs is lack of communication among participating members in the surgery, [35] which may explain our findings that the number of staff had an increasing/decreasing effect on NE occurrence.

A possible limitation of our study is the small number of NEs analyzed when compared to the number of analyzed observations. Therefore, the feature impact estimations should be considered carefully. We used grounded statistical techniques to counter this limitation and plan to further strengthen the results in the future with newly obtained data. In addition, we plan to use transfer learning of NEs from other countries which could be used to better inform our model. [9] The results suggest that the existing “one size fits all” safety approach currently in place may significantly benefit from tailored adjustments that will consider additional factors such as those identified in this work. These more specific guidelines may mitigate a large part of the NE risk and adjust specific risk management programs to improve patient safety.

# 

# **References**

1. Kjellberg J, Wolf RT, Kruse M, et al. Costs associated with adverse events among acute patients. *BMC Health Serv Res* 2017;17(1):1-7.
2. Robert MC, Choi CJ, Shapiro FE, et al. Avoidance of serious medical errors in refractive surgery using a custom preoperative checklist. *J Cataract Refract Surg* 2015;41(10):2171-8.
3. Provisional publication of Never Events reported as occurring between 1 February and 31 March 2018. London, England: National Health Service, April 27, 2018.(https://improvement.nhs.uk/documents/2613/Never\_Events\_1\_February\_to\_31\_March\_2018\_.pdf)
4. Provisional publication of Never Events reported as occurring between 1 April

2018 and 31 January 2019. London, England: National Health Service, February 27, 2019. (https://improvement.nhs.uk/documents/4872/Provisional\_publication\_-\_\_NE\_1\_April\_to\_31\_Jan\_FINAL.pdf)

1. Jung JJ, Jüni P, Lebovic G, et al. First year analysis of the Operating Room Black Box Study. *Ann Surg* 2020;271(1):122-7.
2. Singer SJ, Molina G, Li Z, et al. Relationship between operating room teamwork, contextual factors, and safety checklist performance. *J Am Coll Surg* 2016;223(4):568-80.
3. Surgical Safety Checklist. The World Health Organization, January, 2009. (https://apps.who.int/iris/bitstream/handle/10665/44186/9789241598590\_eng\_Checklist.pdf;jsessionid=1908B5C90ED0DC4F1362F25B6DE63AEA?sequence)
4. Treadwell JR, Lucas S, Tsou AY. Surgical checklists: a systematic review of impacts and implementation*. BMJ Qual Saf* 2014;23(4):299-318.
5. Stawicki SP, Evans DC, Cipolla J, et al. Retained surgical foreign bodies: a comprehensive review of risks and preventive strategies. *Scand J Surg* 2009;98(1):8-17.
6. OECD (2019). Foreign body left in during procedure, 2017 (or nearest year), in Quality and outcomes of care, OECD Publishing, Paris, https://doi.org/10.1787/c37926e7-en.
7. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. *Value Health* 2019;22(7):808-15.
8. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge, England: Cambridge University Press; 2014
9. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach learn* 2006;63(1):3-42.
10. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak* 2011;11:51 https://doi.org/10.1186/1472-6947-11-51
11. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med Res Methodol* 2020;20(1):1-4.
12. Recommended practices for sponge, sharp, and instrument counts. Perioperative Standards and Recommended Practices. Denver, CO AORN. 2010; 207-16.
13. Rodziewicz TL, Houseman B, Hipskind JE. Medical error prevention. Treasure Island, FL Stat Pearls Publishing: 2020.
14. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA*. 2014;311(24):2479-80.
15. Alhusseini MI, Abuzaid F, Rogers AJ, et al. Machine learning to classify intracardiac electrical patterns during atrial fibrillation: machine learning of atrial fibrillation. *Circ Arrhythm Electrophysiol* 2020;13(8): e008160.
16. Gong J, Simon GE, Liu S. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PloS one.* 2019;14(9): e0222665.
17. Moshtaghi O, Haidar YM, Sahyouni R, et al. Wrong-site surgery in California, 2007-2014. *Otolaryngol Head Neck Surg* 2017;157(1):48-52.
18. Steelman VM, Shaw C, Shine L, et al. Unintentionally retained foreign objects: a descriptive study of 308 sentinel events and contributing factors. *Jt Comm J Qual Patient Saf* 2019;45(4):249-58.
19. Koleva SI. A literature review exploring common factors contributing to Never Events in surgery. *J Perioper Pract* 2020;30(9):256-264. doi:10.1177/1750458919886182
20. Thiels CA, Lal TM, Nienow JM, et al. Surgical never events and contributing human factors. *Surgery* 2015;158(2):515-21. doi: 10.1016/j.surg.2015.03.053.
21. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 2009; 338:b2393.
22. Elkan, Charles. "The foundations of cost-sensitive learning." In International joint conference on artificial intelligence, 17, (1), 973-8. Lawrence Erlbaum Associates Ltd, 2001.
23. Nembrini S, König IR, Wright MN. The revival of the Gini importance? *Bioinformatics*. 2018;34(21):3711-8.
24. Hempel S, Maggard-Gibbons M, Nguyen DK, et al. Wrong-site surgery, retained surgical items, and surgical fires: A systematic review of surgical never events. *JAMA surg* 2015;150(8):796-805.
25. Stawicki SP, Moffatt-Bruce SD, Ahmed HM, et al. *J Am Coll Surg* 2013; 216(1):15-22.
26. Freitas PS, Silveira RC, Clark AM, et al. Surgical count process for prevention of retained surgical items: An integrative review. *J Clin Nurs* 2016;25(13-14):1835-47.
27. Gadelkareem RA. Experience of a tertiary-level urology center in the clinical urological events of rare and very rare incidence. I. Surgical never events: 2. Intracorporeally-retained urological surgical items. *Curr Urol* 2017;11(3):151-6.
28. Tofte JN, Caldwell LS. Detection of retained foreign objects in upper extremity surgical procedures with incisions of two centimeters or smaller. *Iowa Orthop J* 2017;37:189.
29. Yoo TK, Oh E, Kim HK, et al. Deep learning-based smart speaker to confirm surgical sites for cataract surgeries: A pilot study. *PloS one* 2020;15(4): e0231322.
30. Elsey EJ, West J, Griffiths G, et al. Time out of general surgery specialty training in the UK: a national database study. *J Surg Educ* 2019;76(1):55-64.
31. Barimani B, Ahangar P, Nandra R, et al. The WHO surgical safety checklist: a review of outcomes and implementation strategies. perioperative care and operating room management. 2020;21: 100117.

Table 1: Characteristics of the data set according to surgical specialty

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Observations n=9234** | | | | | | | **Never Events n=101** |
| Phase  Specialty | \*Pre procedure (n=1,539) (missing data on 760 cases) | Sign in  (n=1,504) | Time out (n=1,498) | First count (n=1,518) | Second count (n=1,501) | Third count  (n=1,498) |  |
| Urology | 72 | 156 | 148 | 124 | 118 | 124 | 7 (6.93%) |
| Orthopedics | 185 | 331 | 324 | 341 | 302 | 326 | 16 (15.84%) |
| ENT | 64 | 105 | 105 | 99 | 102 | 93 | 3 (2.97%) |
| Gynecology | 63 | 143 | 139 | 149 | 153 | 153 | 17 (16.83%) |
| General surgery | 313 | 537 | 558 | 576 | 623 | 604 | 19 (18.81%) |
| Plastic surgery | 22 | 39 | 37 | 40 | 36 | 42 | 2 (1.98%) |
| Vascular surgery | 18 | 45 | 42 | 45 | 42 | 43 | 5 (4.95%) |
| Neurosurgery | 7 | 25 | 19 | 22 | 19 | 19 | 5 (4.95%) |
| Dermatology | 7 | 16 | 26 | 21 | 22 | 24 | 2 (1.98%) |
| Ophthalmology | 12 | 41 | 34 | 33 | 19 | 18 | 8 (7.92%) |
| Maxillofacial | 3 | 12 | 10 | 8 | 10 | 11 | 2 (1.98%) |
| Cardiac and Cardiothoracic | 13 | 54 | 56 | 60 | 55 | 41 | 15 (14.85%) |

Table 2: Characteristics of patients and surgery in the dataset

|  |  |  |
| --- | --- | --- |
| Characteristic | Observations | Never Events |
| Average age | 50.8 years (SD 20.4) | 46 |
| Gender | Male (n=388 (49.8%)), Female (n=391 (50.2%)) | Male (n=46 (45.5%))  Female n=55 (54.5%) |
| Length of surgery | Up to 1 hour: 2124 (23%)  1–2 hours: 4340 (47%)  3–4 hours: 2031 (22%)  Over 4 hours: 739 (8%) | Length of surgery: Up to 1 hour: 54 (53.5%)  1–2 hours: 13 (12.9%)  3–4 hours: 17 (16.8%)  Over 4 hours: 17 (16.8%) |

Table 3: Top 15 contributing features for the six examined departments



