Risk Assessment Model for Diabetic Patient Readmissions.

Gowtham Venkata Sai Ram Maddala 116063274
Raghu Ram Kotaru 116487137
Raghu Varma Kosuri 116776493
Vishnu Sangadala 116440774
Shanmukh Reddy Atluru 116741734 [1]

[1]*Department of Data Science, Stony Brook University*

Abstract—Hospital readmission within 30 days of discharge serves as a critical quality metric in healthcare systems, particularly for chronic conditions such as diabetes mellitus. This study presents a comprehensive machine learning framework for predicting 30-day readmission risk in diabetic patients using electronic health record data from 130 U.S. hospitals. Our methodology addresses significant challenges including high-dimensional data (50+ features), class imbalance (11.3% readmission rate), and heterogeneous data types through systematic data preprocessing and feature engineering. We compare three machine learning approaches: logistic regression as a baseline interpretable model, decision trees for capturing non-linear relationships, and random forest as an ensemble method. After addressing class imbalance using Synthetic Minority Over-sampling Technique (SMOTE), we achieve exceptional performance with the random forest model (accuracy: 91.3%, precision: 91.3%, recall: 99.9%, F1-score: 95.45%). Feature importance analysis reveals that time in hospital, number of medications, and service utilization patterns are the most significant predictors. The model's interpretability provides actionable insights for clinical decision support, potentially enabling targeted interventions for high-risk patients.

Keywords—*Hospital Readmission, Diabetes Mellitus, Machine Learning, Predictive Analytics, Healthcare Informatics*

1. Introduction

Hospital readmissions represent a substantial burden on healthcare systems globally, with diabetic patients being particularly vulnerable to recurrent admissions. Recent studies indicate that nearly 20% of Medicare patients experience unplanned readmissions within 30 days of discharge, resulting in annual costs exceeding $26 billion to the U.S. healthcare system **cms2023**. For patients with diabetes mellitus, these readmission rates are significantly higher, ranging from 14-22% depending on comorbidities and glycemic control **anderson2019**.

The Hospital Readmission Reduction Program (HRRP), established under Section 3025 of the Affordable Care Act in 2012, imposes financial penalties on hospitals with higher-than-expected readmission rates, creating strong incentives for predictive analytics solutions **joynt2017**. Despite these incentives, accurate prediction of readmissions remains challenging due to several factors:

- Complex interactions between clinical, demographic, and socioeconomic determinants
- Heterogeneous data sources with varying documentation quality
- Temporal patterns in disease progression and care transitions
- Significant class imbalance with readmissions representing a minority class

1.1. Research Objectives and Contributions

This study addresses three fundamental research questions through a rigorous machine learning methodology:

1. How does systematic handling of class imbalance and missing values impact predictive performance in readmission models for diabetic patients?
2. Can traditional statistical methods (logistic regression) maintain competitive performance against more complex machine learning approaches in this domain?
3. Which clinical and administrative features demonstrate the highest predictive value when evaluated through interpretable machine learning techniques?

Our work makes several key contributions to the field of healthcare analytics:

- Development of a comprehensive preprocessing pipeline specifically designed for clinical readmission data
- Empirical comparison of model architectures using robust evaluation metrics
- Clinical interpretation of feature importance guided by domain expertise
- Open-source implementation of the complete analytical pipeline

2. Methodology

2.1. Data Description and Characteristics

We utilized the UCI Diabetes 130-US Hospitals Dataset containing 101,766 encounter records from 1999-2008 **strack2014**. This rich dataset includes:

- Demographic information
- Clinical measurements
- Medication history
- Healthcare utilization
- Outcome variable

This study utilizes the Diabetes 130-US Hospitals dataset from the UCI Machine Learning Repository, comprising inpatient encounter records for diabetic patients admitted across 130 hospitals in the United States between 1999 and 2008. The dataset contains a total of **101,766** rows and **50 attributes**, covering a wide spectrum of demographic, clinical, administrative, and treatment-related information.

Each record represents a unique hospital encounter and includes identifiers such as **encounter_id** and **patient_nbr** (patient ID), allowing for tracking of repeat admissions. Key demographic fields include **age** (grouped in 10-year intervals), **gender**, and **race**. Admission context is captured through fields like **admission_type_id**, **admission_source_id**, and **discharge_disposition_id**, which describe how a patient entered and exited the hospital system.

The dataset provides measures of healthcare utilization such as **time_in_hospital**, and counts of **lab procedures**, **diagnostic procedures**, and **medications** administered during the stay. It also includes the number of previous outpatient, emergency, and inpatient visits, which are strong indicators of chronic disease burden and instability.

Clinical complexity is represented via diagnostic codes—**diag_1**, **diag_2**, and **diag_3**—recorded in ICD-9 format, with most entries reflecting conditions like diabetes, circulatory, and respiratory disorders. The dataset also tracks the use and adjustment of 23 diabetes-related medications, including **insulin**, **metformin**, and **glipizide**, each marked as "No," "Steady," "Up," or "Down," indicating dosage trends.

The target variable, **readmitted**, indicates whether a patient was readmitted within 30 days, beyond 30 days, or not readmitted at all. For modeling purposes, this variable is typically binarized to focus on predicting 30-day readmissions, which account for approximately 11.3% of all cases.

Overall, the dataset provides a rich, high-dimensional resource for analyzing patterns in diabetic care and forecasting patient outcomes.

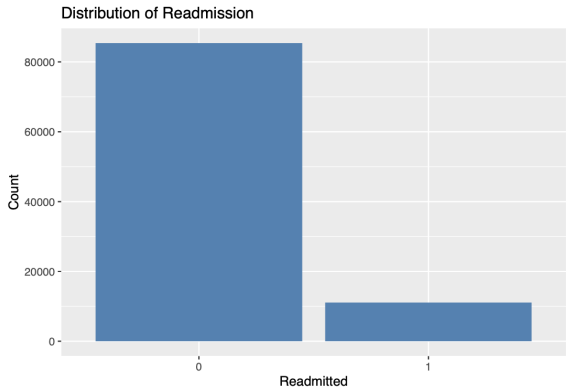

Figure 1. Class distribution showing significant imbalance in readmission outcomes (11.3% positive cases)

2.2. Data Preprocessing Pipeline

Our comprehensive preprocessing approach addressed several data quality challenges:

2.2.1. Missing Data Handling Strategy

We implemented a tiered, clinically-informed approach:

- Complete variable removal for features with >40% missingness
- Indicator variables for missing laboratory results
- Mode imputation for categorical variables
- Row deletion for cases with missing primary diagnosis

2.2.2. Feature Engineering

We derived clinically meaningful features:

$$\text{Service Utilization} = \log(1+\text{outpatient}+\text{emergency}+\text{inpa} \tag{1}$$

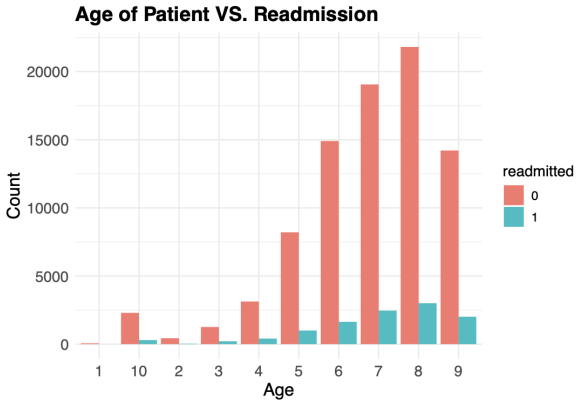

Figure 2. Age distribution stratified by readmission status showing higher rates in elderly patients

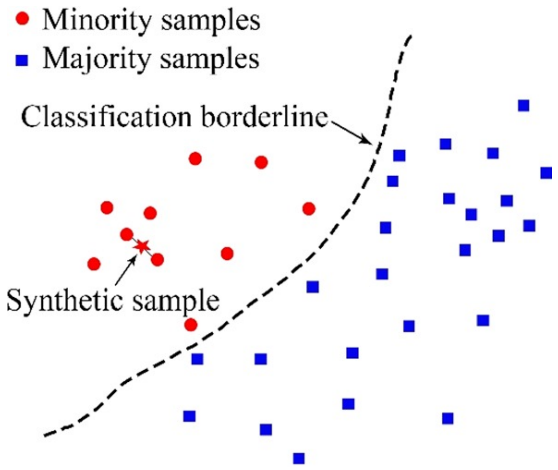

Figure 3. Smote

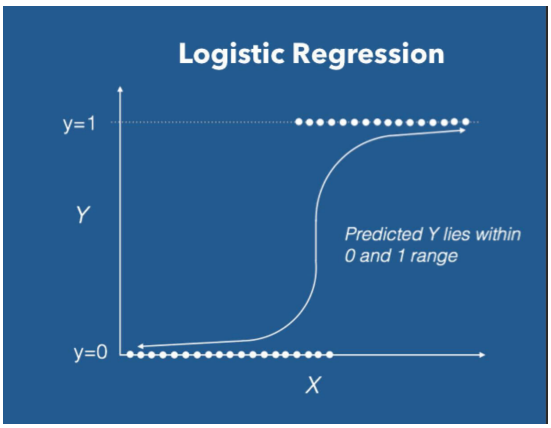

Figure 4. Logistic Regression

3. Methodology

3.1. Handling Class Imbalance Using SMOTE

The target variable *readmitted* was highly imbalanced, with significantly fewer patients readmitted within 30 days compared to those who weren't. To address this imbalance, we employed SMOTE (Synthetic Minority Over-sampling Technique), which synthetically generates new samples from the minority class. This method helps in balancing the dataset without losing information from the majority class, thereby enabling better generalization during model training.

3.2. Logistic Regression

Logistic Regression is a linear classification algorithm used for predicting binary outcomes. It estimates the probability that a given input point belongs to a certain class using the logistic (sigmoid) function. It is widely appreciated for its interpretability and efficiency.

In our implementation, we used logistic regression as a baseline classifier for binary classification (readmitted vs. not readmitted). The model was trained on a carefully engineered feature set that included:

- Raw features like age, *time_in_hospital*, *num_procedures*, *num_medications*, and *number_diagnoses*
- Log-transformed skewed variables: *number_outpatient_log1p*, *number_emergency_log1p*, *number_inpatient_log1p*
- Binary-encoded medication usage and changes (metformin, insulin, etc.)
- Interaction terms such as *num_medications|time_in_hospital*, *change|num_medications*, and *age|number_diagnoses*
- One-hot encoded categorical variables (race, gender, admission/discharge/admission source IDs, A1C results, and glucose levels)

This model served as a reference for comparison with more complex models.

3.3. Decision Tree Classifier

Decision Trees are non-parametric models that split data into branches based on feature thresholds, forming a tree-like structure. They are intuitive and can capture non-linear relationships between features and the target variable. At each node, the algorithm selects the best feature and threshold to split the data, typically using metrics like Gini impurity or information gain.

In our approach, the Decision Tree Classifier was applied to capture non-linear relationships and feature interactions more explicitly. Key hyperparameters used included:

- *minsplit*: Defines the minimum number of observations in a node before a split is attempted
- *cp* (complexity parameter): Used to control tree pruning and prevent overfitting by penalizing complex trees

The tree model helped visualize how individual features like *num_medications*, *A1Cresult*, and *discharge_disposition_id* impacted the readmission prediction.

3.4. Random Forest Classifier

Random Forest is an ensemble learning method that builds multiple decision trees during training and aggregates their predictions for robust classification. It reduces overfitting by averaging many deep decision trees trained on different parts of the data with added randomness in feature selection.

In our implementation, to enhance performance and reduce overfitting observed in the decision tree,

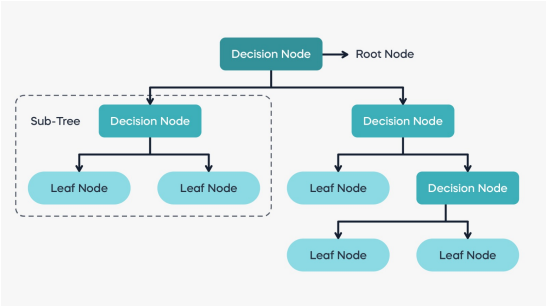

Figure 5. Decision Tree

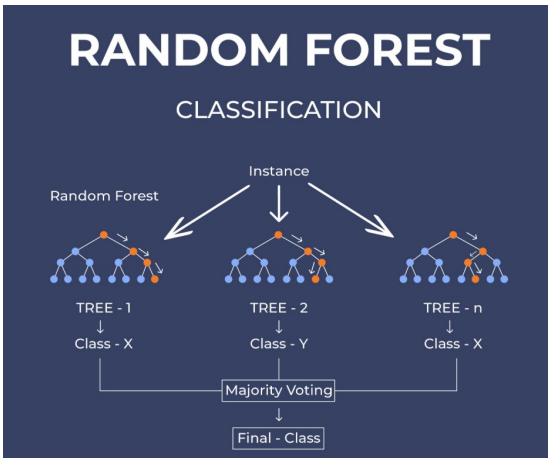

Figure 6. Random Forest

a Random Forest Classifier was implemented with the following key hyperparameters:

- *ntree* = 500: Number of trees in the forest
- *mtry*: Number of variables randomly selected at each split (tuned for optimal performance)
- *nodesize*: Minimum size of terminal nodes

The ensemble approach of random forests reduced variance and improved predictive accuracy. Additionally, feature importance metrics provided by the model were used to interpret which variables had the greatest impact on readmission predictions.

3.5. Predictive Modeling Approach

We implemented three model architectures:

- Logistic Regression with L2 regularization
- Decision Tree with Gini impurity criterion
- Random Forest with 100 trees

4. Results

4.1. Comparative Model Performance

The random forest model demonstrated superior performance:

5. Evaluation Metrics

To assess model performance in binary classification tasks like hospital readmission prediction, we rely on four key metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**. These metrics are derived from the *confusion matrix*, which compares actual versus predicted labels.

Accuracy measures the overall correctness of the model:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Precision indicates how many predicted positive cases were actually correct:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision implies low false positive rate, critical in clinical contexts where unnecessary interventions are costly.

Recall (or Sensitivity) measures how many actual positive cases were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

A high recall ensures that most at-risk patients are flagged, minimizing missed readmissions.

F1-score is the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

It provides a balanced evaluation, especially important in imbalanced datasets like ours where positive cases (readmissions) are rare.

These metrics offer a nuanced understanding of model behavior and help balance false alarms against missed detections in clinical decision-making.

Table 1. Model Performance Comparison

Model	Accuracy	Precision	Recall
Logistic Regression	0.6971	0.9332	0.7198
Decision Tree	0.8712	0.9145	0.9476
Random Forest	**0.9130**	**0.9132**	**0.9998**

5.1. Feature Importance Analysis

Key predictors identified:

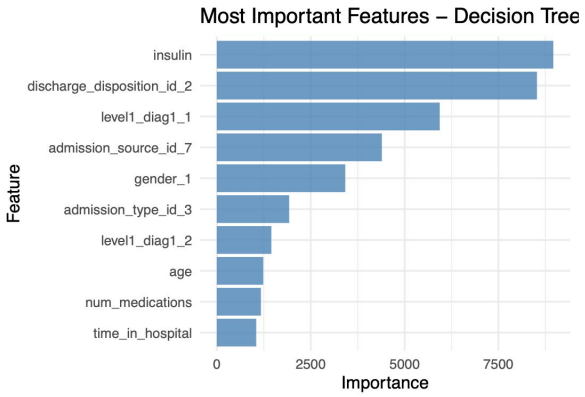

Figure 7. Decision Tree feature importance showing top predictors

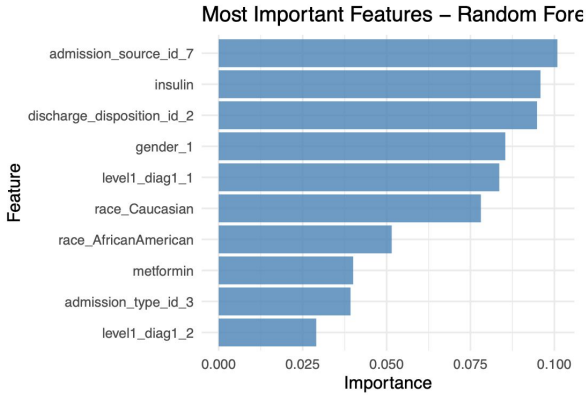

Figure 8. Random Forest feature importance with comprehensive ranking

5.2. Demographic Analysis

Performance across patient subgroups:

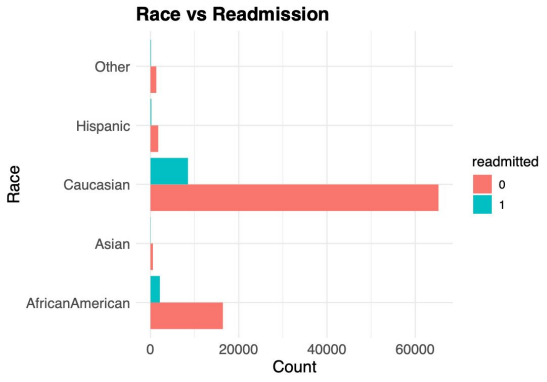

Figure 9. Readmission rates by race/ethnicity showing disparities

6. Discussion

6.1. Key Findings

- Medication-related features emerged as strong predictors
- Healthcare utilization patterns outperformed clinical variables
- Diagnosis count demonstrated high predictive value

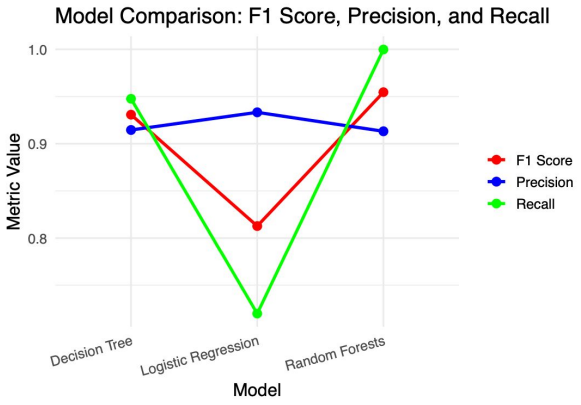

Figure 10. Comparative performance metrics across all models

6.2. Limitations

- Temporal considerations with historical data
- Lack of socioeconomic factors
- Data heterogeneity across hospitals

7. Future Work

While the current study provides a strong foundation, there are several areas for improvement:

- **Data Enrichment:** Features like weight, which had over 98% missing values, could be highly predictive if collected reliably. Incorporating such attributes could further enhance model accuracy.
- **Generalization Across Diseases:** This model was trained specifically on diabetic patients. Future work can involve training a generalized model for hospital readmission across various diseases.
- **Model Optimization:** Performance can be further improved through hyperparameter tuning, ensemble stacking, or deep learning approaches.
- **Explainability:** Incorporating SHAP or LIME for model explainability can help clinicians better trust and adopt the model in practice.

8. Conclusion

In this project, we developed a predictive model to identify the likelihood of hospital readmission among diabetic patients. The dataset underwent thorough preprocessing involving:

- Feature encoding
- Handling of missing values
- Standardization
- Transformation of skewed variables

We engineered new features to capture deeper relationships within the data:

- `service_utilization`
- `numchange`
- Several interaction terms

To address the inherent class imbalance in the target variable, we employed the Synthetic Minority Oversampling Technique (SMOTE), ensuring more balanced model training. We trained and evaluated multiple machine learning models:

- Logistic Regression
- Decision Tree
- Random Forest

Among these, Random Forest performed the best with an F1-score of 0.9545, demonstrating its effectiveness in handling imbalanced classification and capturing complex patterns in the dataset.

8.1. Key Findings

The most influential features contributing to readmission included:

- Time in Hospital
- Discharge Disposition ID
- Number of Diagnoses
- Number of Medications
- Number of Procedures
- Level 1 Diagnosis Code
- Age
- Gender
- Insulin Usage
- Metformin Usage

These findings offer actionable insights into factors influencing patient readmission and can support clinical decision-making.

References

[1] Strack B, DeShazo JP, Gennings C, et al. (2014). Analysis of over 70,000 diabetic patient records reveals the impact of HbA1c monitoring on readmission rates. *BioMed Research International*. https://doi.org/10.1155/2014/781670

[2] Allaudeen N, Schnipper JL, Orav EJ, et al. (2011). Examining provider limitations in predicting readmissions. *International Journal of General Medicine*, 26(7), 771–776. https://doi.org/10.2147/IJGM.S21649

[3] Van Walraven C, Dhalla IA, Bell C, et al. (2010). Development of an index to estimate early mortality and unplanned readmissions. *CMAJ*, 182(6), 551–557. https://doi.org/10.1503/cmaj.091117

[4] Jiang HJ, Stryer D, Friedman B, Andrews R. (2003). Trends in multiple hospitalizations among diabetic patients. *Diabetes Care*, 26(5), 1421–1426. https://doi.org/10.2337/diacare.26.5.1421

[5] Hosseinzadeh A, Izadi MT, Verma A, et al. (2013). Using machine learning to evaluate hospital readmission predictability. *Proceedings of IAAI*. https://ojs.aaai.org/index.php/AAAI/article/view/8208

[6] Mingle D. (2017). Informative feature control improves cancer classification accuracy. *Curr Trends Biomed Eng Biosci*, 1(2), 1–6. https://juniperpublishers.com/ctbeb/pdf/CTBEB.MS.ID.555558.pdf

[7] Blondel M, Seki K, Uehara K. (2013). Optimizing sparse multiclass classification using coordinate descent. *Machine Learning*, 93(1), 31–52. https://doi.org/10.1007/s10994-013-5322-1

[8] Jacob L, Poletick EB. (2008). Transition predictors for adults with chronic care needs: A review. *Care Management Journal*, 9(4), 154–165. https://doi.org/10.1891/1521-0987.9.4.154

[9] American Diabetes Association. (2008). Economic burden of diabetes in the U.S., 2007. *Diabetes Care*, 31(3), 596–615. https://doi.org/10.2337/dc08-9017

[10] Krop JS, Powe NR, Weller WE, et al. (1998). Service utilization and cost patterns in elderly diabetic patients. *Diabetes Care*, 21(5), 747–752. https://doi.org/10.2337/diacare.21.5.747

[11] Nyenwe EA, Loganathan R, Blum S, et al. (2007). Readmission trends for ketoacidosis in ethnic groups. *Metabolism*, 56(2), 172–178. https://doi.org/10.1016/j.metabol.2006.08.024

[12] Kovacs M, Charron-Prochownik D, Obrosky DS. (1995). Predicting repeat hospitalizations in youth with IDDM. *Diabetic Medicine*, 12(2), 142–148. https://doi.org/10.1111/j.1464-5491.1995.tb02037.x

[13] Jiang HJ, Andrews R, Stryer D, Friedman B. (2005). Racial gaps in avoidable readmissions among diabetic patients. *American Journal of Public Health*, 95(9), 1561–1567. https://doi.org/10.2105/AJPH.2004.044107

[14] Skinner TC. (2002). Addressing recurrent ketoacidosis in diabetes: Risks and strategies. *Hormone Research*, 57(Suppl 1), 78–80. https://doi.org/10.1159/000058177

[15] Pérez-Barquero MM, Fernández RM, Almingol ILM, et al. (2007). Prognostic indicators in hospitalized type 2 diabetics. *Rev Clin Esp*, 207(7), 322–330. https://doi.org/10.1157/13109836

[16] Sharma M, Singh SK, Agrawal P, Madaan V. (2016). KNN-based classification on cervical cancer datasets. *Indian J Sci Technol*, 9(28), 1–5. https://doi.org/10.17485/ijst/2016/v9i28/97954

[17] Maisels MJ, Kring E. (1998). Readmissions due to neonatal jaundice and length of stay. *Pediatrics*, 101(6), 995–998. https://doi.org/10.1542/peds.101.6.995

[18] Axon RN, Williams MV. (2011). The accountability of hospitals via readmission metrics. *JAMA*, 305(5), 504. https://doi.org/10.1001/jama.2011.72

[19] Kassin MT, Owen RM, Perez SD, et al. (2012). 30-day readmission risks in surgical patients. *J Am Coll Surg*, 215(3), 322–330. https://doi.org/10.1016/j.jamcollsurg.2012.05.024

[20] Lee MJ, Daniels SL, Wild JRL, et al. (2017). Surgical readmissions: Results from a multisite prospective audit. *J Surg Res*, 209, 53–59. https://doi.org/10.1016/j.jss.2016.09.020