Interrater Reliability of Student vs. Expert Annotations for Machine Learning in Traumatic Wound Segmentation and Tissue Grading

Daniel C. Thompson, MBBS[1,2]; Matthew Maceda, BS[1]; Scott F. Grey, PhD[1,3]; Felipe A. Lisboa, MD[1,3,4]; Omri Shental, MD[1,5]; Eric A. Elster, MD[1,4]; Seth A. Schobel, PhD[1]

[1] Uniformed Services University (USU) Surgical Critical Care Initiative (SC2i), Bethesda, MD; [2] Royal Centre for Defence Medicine, Birmingham, UK; [3] The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD; [4] Walter Reed National Military Medical Center, Bethesda, MD; [5] Israel Defense Forces Medical Corps, Tel Hashomer, Ramat Gan, Israel

INTRODUCTION

Clinical need:
- Accurate assessment of large traumatic wounds is essential for treatment planning and monitoring healing.
- Current wound assessment relies on visual inspection, which can be subjective and vary between clinicians.

Machine learning for wound assessment:
- Machine learning has the potential to improve the objectivity, consistency, and efficiency of wound assessment.
- Algorithm development requires high-quality **annotated datasets**, where wound images are labeled for clinically relevant features.
- **Segmentation** outlines wound regions, while **tissue grading** scores features such as infection, nonviable tissue, and granulation tissue.

Annotation challenge:
- Expert annotation is labor-intensive and limits dataset size; trained medical students may offer a scalable alternative.

Objective/Hypothesis:
- We hypothesized that protocol-trained medical students would achieve expert-level annotation reliability sufficient for machine-learning development.

METHODS

Data curation:
- A standardized traumatic wound image dataset was assembled from two observational cohorts:
  - **WoundVAC** (combat casualties; USU IRB #352334, 2007–2012)
  - **WounDx** (civilian trauma patients; Grady IRB #00058229, 2012–2015)
- Images spanned diverse wound types and healing outcomes during serial debridements prior to delayed closure.
- 100 WoundVAC images were selected for semantic segmentation of the wound bed, surrounding skin, and reference ruler; a subset (N=45) underwent tissue composition grading.

Student training:
- Second- and third-year medical students served as annotators.
- Their training included a literature review, instructor-led sessions on traumatic wound care and image interpretation, guided annotation exercises, and scoring practice using standardized examples.
- Feedback from expert annotators (physicians who developed the segmentation and grading protocols) was provided throughout training.
- Students and experts independently evaluated wound images using the same annotation procedures.

Wound segmentation:
- Annotators delineate boundaries on each image "mask."
- Each image received three paired masks from both a student and an expert (wound, skin, ruler).

Original Image | Student Mask | Expert Mask

Figure 1: Original wound image and student and expert wound bed masks.

- Masks are transformed into matrices that label each pixel either a 0 or 1.
- **Dice Similarity Coefficient (DSC)** ranging from 0 to 1, was used to assess interrater reliability by measuring spatial overlap between paired matrices, with 1 indicating perfect overlap.
- Mean DSCs were tested against a predefined reliability threshold of 0.80 using one-sided t-tests with Benjamini–Hochberg FDR adjustment. Comparisons across groups (mask types, raters, wound types, and outcomes) were assessed with Kruskal–Wallis tests.

Tissue composition grading:
- Students and experts graded wound characteristics using the ordinal scales in Table 1.
- Interrater reliability was evaluated using **Gwet's Agreement Coefficient (AC2),** with one-sided Z-tests against a 0.80 threshold (Benjamini–Hochberg FDR-adjusted) and ANOVA-like bootstrap tests for group comparisons (tissue types, raters, wound types, study, and outcomes).

Table 1: Standardized tissue composition grading criteria.

Grade	Signs of infection	Non-viable tissue	Granulation tissue
0	Absence of signs of infection	Absence of non-viable tissue	Absence of granulation tissue
1	Minor, some, or isolated infection / inflammation (including purulence)	Minimal non-viable tissue in < 3 isolated sites in < 10% of the wound surface	Isolated granulation tissue sites in < 10% of the wound surface
2	Gross or major infection / inflammation (including purulence)	Moderate non-viable tissue in between 3 to 5 isolated sites, in 10 to 20% of the wound surface	Moderate granulation tissue in < 50% of the wound surface
3		Major non-viable tissue > 5 isolated sites, or major necrosis area in >20% of the wound surface	Major granulation tissue in > 50% of the wound surface

RESULTS

Table 2: DSC for image segmentation across experts and students

Group	N Rated	Mean DSC* (95% CI)	Adjusted p-value†	Omnibus p-value‡
Mask Type:				**<0.001**
Ruler	90	0.98 (0.98, 0.98)	<0.001	
Skin	99	0.96 (0.94, 0.97)	<0.001	
Wound	100	0.98 (0.98, 0.98)	<0.001	
Wound Type:				**0.596**
Amputation	6	0.89 (0.65, 1.00)	0.190	
Fasciotomy	82	0.98 (0.97, 0.98)	<0.001	
Open Fracture	62	0.97 (0.97, 0.98)	<0.001	
STI	136	0.97 (0.97, 0.98)	<0.001	
Unknown	3	0.97 (0.93, 1.00)	0.002	
Outcome:				**0.483**
Dehisced	42	0.97 (0.97, 0.98)	<0.001	
Delayed	29	0.97 (0.96, 0.98)	<0.001	
Healed	215	0.97 (0.97, 0.98)	<0.001	
Unknown	3	0.97 (0.93, 1.00)	0.002	
Rater:				**0.278**
Student 1	143	0.97 (0.96, 0.98)	<0.001	
Student 2	146	0.98 (0.97, 0.98)	<0.001	

Wound segmentation results:
- Student and expert segmentations showed excellent agreement.
- Skin masks had slightly lower agreement than wound and ruler masks but remained highly reliable.
- Agreement was consistent by wound type, outcome, or student rater.

* Subgroup mean of the DSC of pixel-wise agreement between paired segmentations.
† Adjusted p-values are one-sided and test the hypothesis that the mean Dice score is above the minimum acceptable value of 0.80.
‡ Omnibus p-values are derived from Kruskal–Wallis tests.

Table 3: AC2 for grading wound severity across experts and students

Group	N Rated	Gwet AC2* (95% CI)	Adjusted p-value†	Omnibus p-value‡
Score Type:				**0.038**
Closure	42	0.88 (0.74, 1.00)	0.136	
Granulation	42	0.89 (0.73, 1.00)	0.136	
Infection	42	0.96 (0.91, 1.00)	<0.001	
Nonviable	42	0.84 (0.75, 0.93)	0.181	
Wound Type:				**0.450**
Amputation	48	0.97 (0.95, 0.99)	<0.001	
Fasciotomy	8	0.94 (0.84, 1.00)	<0.001	
Open Fracture	16	0.91 (0.81, 1.00)	0.011	
STI	80	0.94 (0.89, 0.99)	<0.001	
Outcome:				**0.886**
Dehisced	24	0.96 (0.93, 1.00)	<0.001	
Delayed	44	0.96 (0.91, 1.00)	<0.001	
Healed	84	0.95 (0.90, 0.99)	<0.001	
Rater:				**0.101**
Student 1	120	0.97 (0.95, 0.98)	<0.001	
Student 2	32	0.89 (0.76, 1.00)	0.102	
Study:				**0.516**
WounDx	52	0.97 (0.94, 0.99)	<0.001	
WoundVAC	116	0.95 (0.91 ,0.98)	<0.001	

Tissue composition grading results:
- Student and expert severity grades showed excellent agreement.
- Agreement was highest for infection grading and lowest for nonviable tissue grading.
- Agreement was consistent across wound types, outcomes, raters, and studies.

* AC2 is Gwet's agreement coefficient for ordinal scores.
† Adjusted p-values are one-sided and test the hypothesis that the Gwet AC2 agreement statistic is above the minimum acceptable value of 0.80.
‡ Omnibus p-values are derived from ANOVA-like bootstrap tests.

CONCLUSION

- Medical students trained under standardized protocols produced annotations closely matched by expert physicians.
- This model exposes students to machine learning development while expanding the pool of skilled annotators, functioning as a workforce multiplier for scalable dataset generation in automated wound assessment.

*SC2i is a consortium of Uniformed Services University of the Health Sciences, Walter Reed National Military Medical Center, Duke Health, Emory University School of Medicine, and Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc.