Predicting Hemorrhage related Outcomes with CT Volumetry for Traumatic Hemothorax

Benjamin Albert, Gary Yang, Chang Yan
Mentors: Professor David Dreizin and Professor Mathias Unberath
The Johns Hopkins University and The University of Maryland School of Medicine

Computer Integrated Surgery II - Teams
EN.601.496

Introduction

Hemothorax (HTX)—blood accumulation in the pleural cavity around the lungs—patients are currently treated by qualitative estimates of blood volume using CT scans. To automate this analysis, deep neural networks are employed to segment hemothoraces from patient CT scans. The network segmentation is converted to an estimated volume, yielding an adjusted R of 0.91 compared with manually segmented volume, which is then used as a predictor for a composite variable: patient requires massive transfusion or dies. Together with high-level patient data, a random forest classifier achieves an auROC of 0.944, indicating strong predictive capabilities for the composite variable.

Problem

Qualitative grading for HTX are
- imprecise and subjective
- reliant on expert radiologists for reliability
- sometimes inconsistent among experts

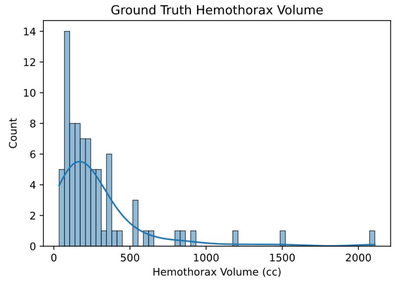
Manual segmentations for HTX are
- precise but time-impermissible

Prior automated models for pleural effusion (excess liquid) are [4]
- rule-based or atlas-based
- insufficient in handling
  - anatomical distortion
  - heterogeneity of attenuation
  - traumatic lung scans

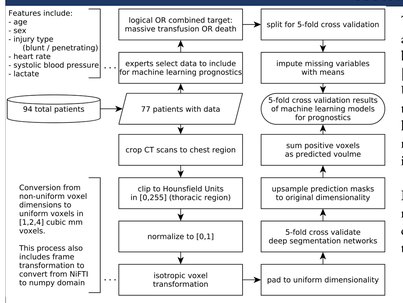
No prior attempt for automatic HTX volumetry, though necessary for it helps doctors with planning.

Data

- 94 cases
  - hemothorax CT scans
  - corresponding manual segmentations
  - baseline demographics and clinical characteristics
- 77 cases remained upon removing
  - CT scans with erroneous voxel dimension metadata
  - Patients without clinical data readings
- 6 clinical variables available at point of care are chosen
  - age, sex, injury type, heart rate, systolic blood pressure, lactate concentration
- Volume estimation + clinical variables are used to predict a composite patient outcome (massive transfusion or mortality)



Methods

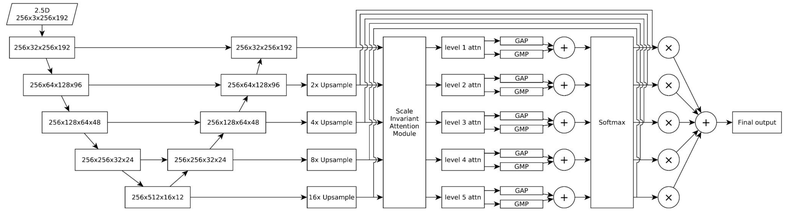


Three deep networks are evaluated: UNet (2.5D) [1], UNet 3D [2], and UNet-FAN. UNet-FAN, the architecture of which is illustrated below, was developed as a combination UNet (2.5D) and PIPO-FAN [3]. PIPO-FAN validation yielded poor performance, so the trained UNet models were used as replacement to the PIPO module to train the FAN scale-invariant attention module post hoc. This transfer learning approach allowed the FAN module to apply attention mechanisms to the multiscale features learned in UNet, slightly improving dice.

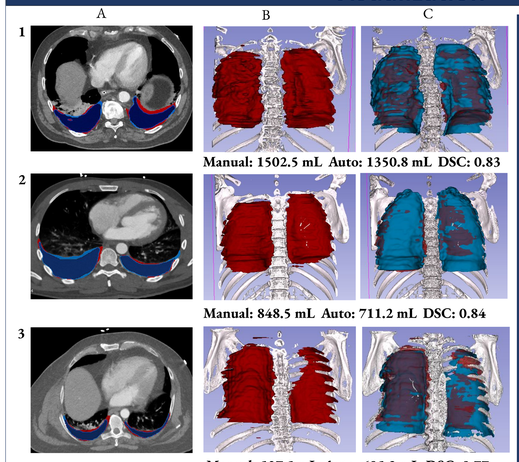
It was observed that training deep segmentation networks on left and right lungs individually yielded superior dice scores than from training on the union of the left and right masks. The final predicted volume takes the union of the left and right prediction masks.

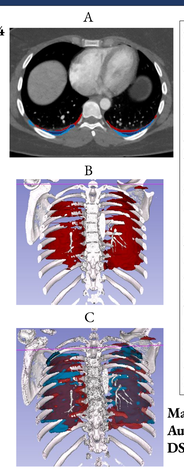
Logistic regression was performed to ascertain the predictive power of HTX volume for the composite variable using 5-fold cross validation. The results of logistic regression are then compared to the logistic regression performance on the manual qualitative estimates from the consensus of two expert radiologists, categorized as low, medium, and high.

In addition to univariate analysis, clinical features are used for multivariate predictions. Eight machine learning models [6] are evaluated: logistic regression, bayesian network with global architecture search, discrete naive bayes, gaussian naive bayes, decision table, linear support vector machine, RBF support vector machine, and random forest. The random forest models performed best, demonstrating comparable performance between the manual and automatic features for the composite variable prognostics.



Visualization





(A) Overlap (purple) of manual and automated (FAN UNET) hemothorax labels on axial images. **(B)** 3D rendering of automated label. **(C)** 3D rendering of overlap (purple) between automated (red) and manual (blue) labels.

Results

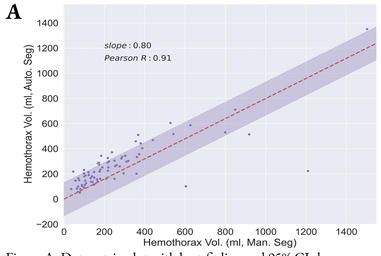


Figure A: Dot matrix plot with best-fit line and 95% CI shows correlation between automated volume (vol.) and manual hemoperitoneum volume. The prediction from human expert and our deep learning is consistent.

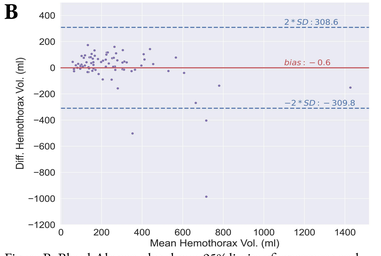
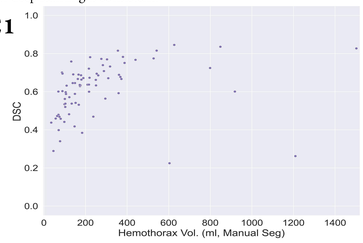


Figure B: Bland-Altman plot shows 95% limits of agreement and measurement bias. On average, there is a 0.6-mL underestimation by the deep learning algorithm. The bias is relatively small and standard deviation is 155.6 mL.



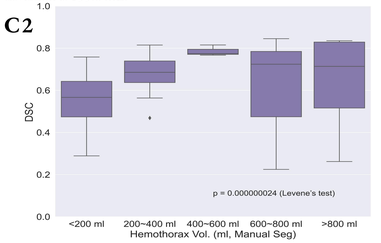


Figure C: Distribution of Dice similarity coefficients (DSCs). The box plot in C2 shows DSC improves/variance decreases with increasing vols at volume range 0-600 ml, (Levene's test, $p < 0.00001$) explaining low DSCs in rows 4 and 5 (image left). In volume range >600ml, we have only 7 instance and some of them are outliers, so the behavior in this range is not clear.

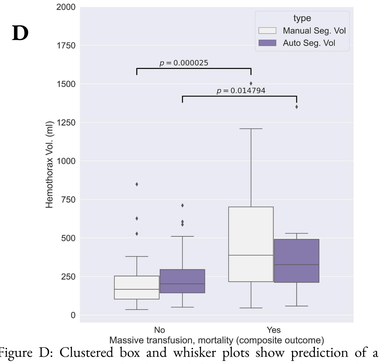


Figure D: Clustered box and whisker plots show prediction of a composite outcome for the need for massive transfusion and in hospital mortality. Manual and HTXvol-auto vols both have significant association with composite outcome (MT + IHM), with $p = 0.0003$ and 0.015 respectively.

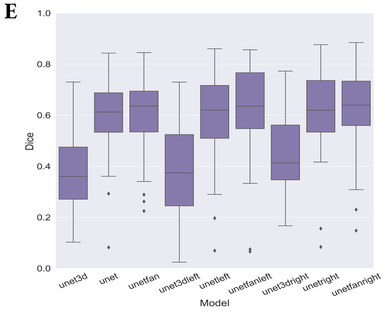


Figure E: Box plot of Dice score of all deep net modes used. UNet-FAN achieves slightly higher dice score than UNet and much better than UNet 3D for all data: left lung, right lung, and the union of the lungs.

Data	Model	MCC	AUROC	AUPRC_N	AUPRC_P	RMSE
uni_qual	Logistic	0.4931	0.7609	0.6219	0.6590	0.5390
uni_ufan	Logistic	0.1746	0.7081	0.7232	0.4900	0.5344
mul_qual	RF	0.6370	0.9450	0.9860	0.8550	0.2829
mul_ufan	RF	0.6870	0.9440	0.9870	0.7680	0.2831

Table: UNet-FAN alone does not predict the composite patient outcome as well as the qualitative measurement. With the addition of clinical data, the automated multivariate model achieves comparable performance to the qualitative manual ones.

Discussion

In general, the deep network predicted volume and manual segmented volume are highly associated with adjusted R=0.91 and the bias is very low at -0.6 mL. The Dice similarity coefficients improves and its variance decrease as volume increase. The small hemothoraces with lesser dice scores are clinically insignificant compared to the larger accumulations of blood. Therefore, it is important that the performance of the automated volume estimates is best for larger hemothoraces.

Both manual and predicted volume have significant association with the require for mass transfusion and in-hospital mortality.

We are able to predict the composite outcome of MT+IHM using automated prediction volume and 6 patient metadata (Age, Sex, HR, BP, lactate, injury-type: blunt / penetrating) with random forest model and reach an auROC of 0.9440. This is at least as good as using expert information from 2 radiologists.

The results suggest that the automated methods can replace expert analysis with comparable performance, thereby reducing costs, labor, and improving availability to accurate prognostics.

Future Work

1. Computing loss for each decoding layer
2. Adding data augmentation to counter the lack of large HTX cases
3. Real-world clinical application

Lessons Learned

1. Time estimates were too short, particularly for the maximal deliverables
2. Agile development was more effective than waterfall methods
3. Concentrating developer time on related tasks was most effective
4. Making scripts/executables flexible with respect to the environment, such as directory structures and command line arguments, was important

Acknowledgements

We thank Bryan Nixon for analyzing the clinical data, Jean Jeudy for consensus volume estimates, Nahye Kim for helping to manually label volumes, Guang Li for image data curation, and Anna Zapaishchykova for code contribution.

References

[1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.

[2] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. InInternational conference on medical image computing and computer-assisted intervention 2016 Oct 17 (pp. 424-432). Springer, Cham.

[3] Fang X, Yan P. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging. 2020 Jun 9;39(11):3619-29.

[4] Sangster, G. P., González-Beicos, A., Carbo, A. I., Heldmann, M. G., Ibrahim, H., Carrascosa, P., Nazar, M., & D'Agostino, H. B. (2007). Blunt traumatic injuries of the lung parenchyma, pleura, thoracic wall, and intrathoracic airways: multidetector computer tomography imaging findings. Emergency radiology, 14(5), 297–310.

[5] Dreizin, D., Zhou, Y., Fu, S., Wang, Y., Li, G., Champ, K., Siegel, E., Wang, Z., Chen, T., & Yuille, A. L. (2020). A Multiscale Deep Learning Method for Quantitative Visualization of Traumatic Hemoperitoneum at CT: Assessment of Feasibility and Comparison with Subjective Categorical Estimation. Radiology. Artificial intelligence, 2(6), e190220.

[6] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009 Nov 16;11(1):10-8.