

Evaluating Estimation Techniques in Medical Imaging Without a Gold Standard: Experimental Validation

John W. Hoppin^a, Matthew A. Kupinski^{b,c}, Donald W. Wilson^b,
Todd Peterson^b, Benjamin Gershman^c, George Kastis^b, Eric Clarkson^{a-c},
Lars Furenlid^{b,c} and Harrison H. Barrett^{a-c}

^aProgram in Applied Mathematics

^bDepartment of Radiology

^cOptical Sciences Center

The University of Arizona, Tucson, AZ

ABSTRACT

Imaging is often used for the purpose of estimating the value of some parameter of interest. For example, a cardiologist may measure the ejection fraction (EF) of the heart to quantify how much blood is being pumped out of the heart on each stroke. In clinical practice, however, it is difficult to evaluate an estimation method because the gold standard is not known, e.g., a cardiologist does not know the true EF of a patient. An estimation method is typically evaluated by plotting its results against the results of another (more accepted) estimation method. This approach results in the use of one set of estimates as the pseudo-gold standard. We have developed a maximum-likelihood approach for comparing different estimation methods to the gold standard without the use of the gold standard. In previous works we have displayed the results of numerous simulation studies indicating the method can precisely and accurately estimate the parameters of a regression line without a gold standard, i.e., without the x-axis. In an attempt to further validate our method we have designed an experiment performing volume estimation using a physical phantom and two imaging systems (SPECT,CT).

Keywords: Regression analysis, image quality, parameter estimation

1. INTRODUCTION

We have previously developed a method for comparing estimation tasks without a gold standard [1,2]. Our method came in response to a need in the medical imaging community for objective comparison of estimation methods performed using different imaging systems. For example, researchers might want to know which imaging modality (ultrasound, MRI, or SPECT) should be used to best estimate an individual's cardiac ejection fraction. Our method is analogous to the techniques initially developed by Henkelman, *et al.* [3], for assessing observer performance on classification tasks without the use of ground truth.

Comparing classification tasks without truth is well studied [3–6], whereas the problem of evaluating estimation tasks without a gold standard has received substantially less attention. Many researchers have attempted to compare estimation tasks by measuring the relationship (via regression analysis) between their estimates and the estimates of a more accepted imaging modality [7–13]. However, since the more accepted modality is rarely considered the gold standard, this type of analysis is faulty. Techniques have been developed [14,15] that attempt to quantify the agreement between the estimates of two imaging modalities. These techniques, however, do not address the relationship between the estimates and the truth.

We have performed extensive studies using simulated data to better understand the performance of our method. These studies have largely been successful, yet received the usual, and justified, skepticism associated with simulation studies. Thus, to address this skepticism we have performed a phantom study involving volume estimation using both computed tomography (CT) and single photon emission computed tomography (SPECT).

Corresponding author: J.W.H., E-mail: jhoppin@math.arizona.edu, Address: Department of Radiology, PO Box 245067, Tucson, AZ 85724-5067

2. METHOD

We present a brief synopsis of our method Regression Without Truth (RWT) developed previously in Hoppin *et al.* [1] and Kupinski *et al.* [2]. We begin with an equation relating the gold standard Θ_p for patient p to the estimate θ_{pm} for patient p using modality m given by,

$$\theta_{pm} = a_m \Theta_p + b_m + \epsilon_{pm}, \quad (1)$$

where a_m and b_m are the linear model parameters and ϵ_{pm} is the noise term. The linear model parameters characterize the mapping of the gold standard to its estimate. These linear model parameters are specific to the modality m and independent of the patient p .

We assume the noise term ϵ_{pm} is Gaussian distributed with mean zero and standard deviation σ_m (another linear model parameter). This assumption yields a Gaussian probability density function (PDF) for the estimates conditional upon the linear model parameters and the gold standard, *i.e.* $pr(\{\theta_{pm}\}|\{a_m\}, \{b_m\}, \{\sigma_m\}, \Theta_p)$. We must now consider a parameterized PDF $pr(\Theta_p|\{\zeta_i\})$ associated with the unknown gold standard, *e.g.*, the population distribution of cardiac ejection fraction, bone density, etc. Using this PDF we marginalize over the unknown gold standard via

$$pr(\{\theta_{pm}\}|\{a_m\}, \{b_m\}, \{\sigma_m\}, \{\zeta_i\}) = \int d\Theta_p pr(\Theta_p|\{\zeta_i\}) pr(\{\theta_{pm}\}|\{a_m\}, \{b_m\}, \{\sigma_m\}, \Theta_p). \quad (2)$$

Note that we have added the list of parameters characterizing the gold standard distribution $\{\zeta_i\}$ to the list of conditional parameters.

In Hoppin *et al.* [1] we derive an expression for the log-likelihood of the unknown parameters ($\{a_m\}, \{b_m\}, \{\sigma_m\}, \{\zeta_i\}$) given the estimates from multiple modalities on a common group of patients. This expression is given by

$$\lambda(\boldsymbol{\eta}|\{\theta_p\}) = -\frac{P}{2} \sum_{m=1}^M \ln(2\pi\sigma_m^2) + \sum_{p=1}^P \ln \left[\int d\Theta_p pr(\Theta_p|\{\zeta_i\}) \exp \left(-\sum_{m=1}^M \frac{1}{2\sigma_m^2} (\theta_{pm} - a_m \Theta_p - b_m)^2 \right) \right], \quad (3)$$

where $\boldsymbol{\eta} = [\{a_m\}, \{b_m\}, \{\sigma_m\}, \{\zeta_i\}]$. We maximize Eq. 3 to produce maximum-likelihood estimates of the linear model parameters as well as the parameters characterizing the gold standard distribution. One can then use the linear model parameters to compare the estimation techniques. Specifically, by solving for the gold standard Θ_p in Eq. 1 we arrive at a random variable with standard deviation σ_m/a_m which we can estimate by $\hat{\sigma}_m/\hat{a}_m$. This quantity serves as a figure of merit in determining which modality returns better estimates of the parameter of interest.

3. DESIGN OF EXPERIMENT

Our experiment to validate RWT consists of estimating multiple volumes in a phantom using a dual-modality (CT/SPECT) imaging system developed by our group [16]. The CT component of the dual-modality system is comprised of an Oxford Instruments (XTF5000/75) x-ray tube and a Kodak KAF-1001E series 1024×1024 pixel CCD array with an active area of $5.0 \times 5.0 \text{cm}^2$. The SPECT component of the dual-modality system consists of the compact Cadmium Zinc Telluride (CdZnTe) semiconductor camera with field of view $2.5 \times 2.5 \text{cm}^2$ developed previously in our group [17]. Note that the tomographic data in both systems are generated by rotating the object rather than the camera. A schematic diagram is given in Fig. 1.

We fabricated the phantom by drilling out an asymmetric pattern in a 2.5cm diameter plexiglass cylinder. A 3D reconstruction of the phantom is given in Fig. 2 (note that the reconstruction is inverted in an attempt to better display the complex nature of the phantom). The phantom has a volume of approximately 4ml. We used a solution consisting of 25% ^{99m}Tc -pertechnetate (typically 8mCi/ml), 5% omnipaque (an x-ray contrast

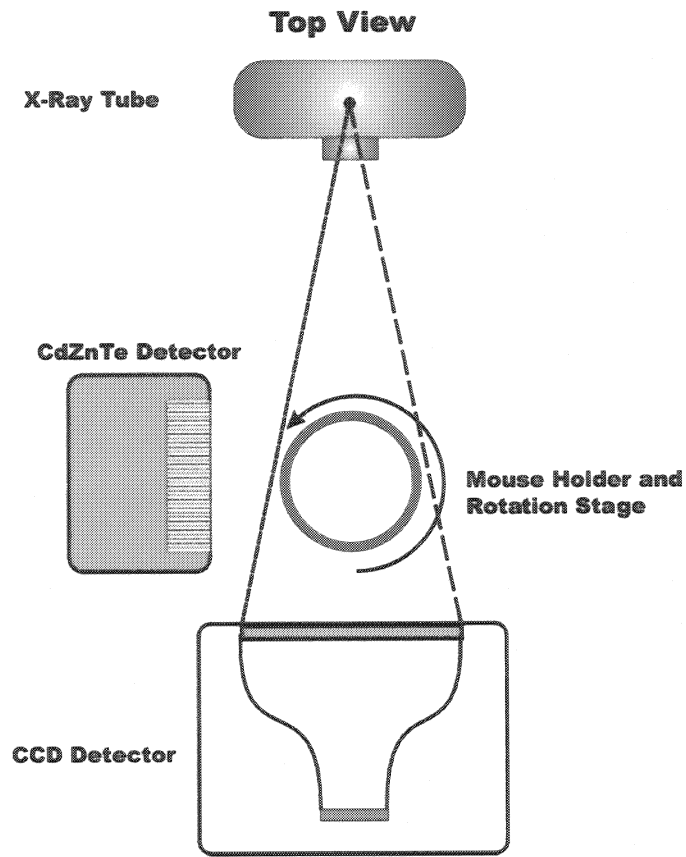


Figure 1: Schematic diagram for the dual-modality imaging system.

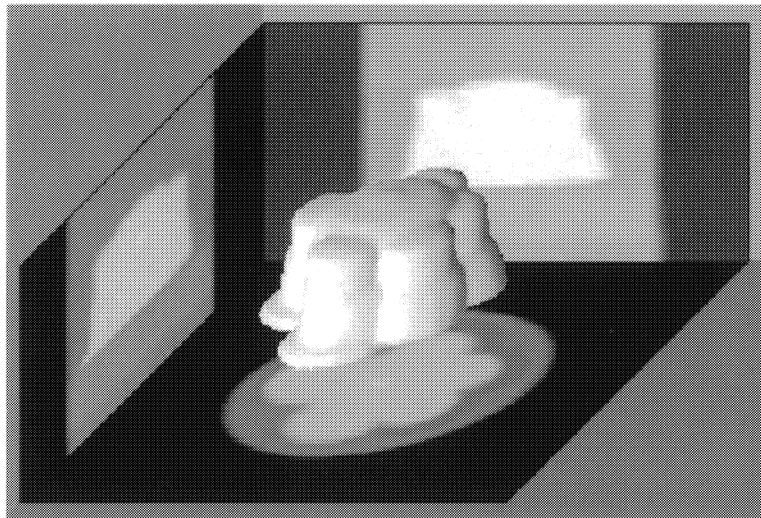


Figure 2. A 3D Reconstruction of the phantom imaged with 3.06ml of solution. Note that the reconstruction is inverted in an attempt to better display the complex nature of the phantom.

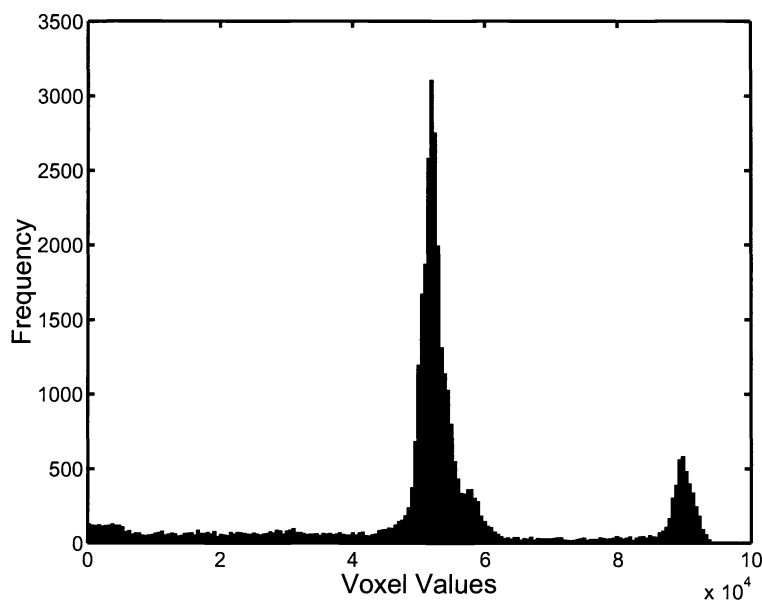


Figure 3. A histogram of positive voxel values from a CT reconstruction of the phantom imaged with 3.06ml of solution. The two peaks consist of voxel values corresponding to plexiglass ($\sim 50,000$) and solution ($\sim 90,000$). A majority of the voxel values corresponding to air are negative, and thus not shown. Note that the voxel values are not given in Hounsfield units.

agent), and 70% water. We imaged 25 volumes with values we sampled from a truncated normal distribution with lower and upper bounds of 0.5 and 3.5ml, respectively. The phantom was filled to the predetermined volumes using a pipette accurate to $\pm 2\mu\text{l}$. Given the accuracy of the pipette, it is a gold standard of volume for this experiment.

Image data were acquired at 180 projection angles with 1 second exposures on the CT system. The SPECT data were taken at 60 projection angles each with 35 seconds of exposure. The data collected using the CT system were reconstructed on a $64 \times 64 \times 32$ voxelized grid, while data collected using the SPECT system were reconstructed on a $64 \times 64 \times 64$ voxelized grid. All data were reconstructed using filtered back projection.

We thresholded voxel values in order to segment out the solution in the image reconstructions. For the SPECT reconstructions we chose our threshold values manually using a gray-level histogram for each image. We generated two sets of volume estimates using the CT data. The first estimation approach, CTI, consisted of manual thresholding and included magnification correction. The second estimation approach, CTII, used a fixed threshold and did not account for magnification. Thus the relationship between the estimates obtained using CTII and the gold standard is quantified with a slope greater than one. In Fig. 3 we display a histogram of voxel values from a CT reconstruction.

4. RESULTS

We applied RWT to the three sets of volume estimates obtained in the experiment resulting in estimated slopes, intercepts, and noise terms relating the volume estimates to the gold standard. This analysis did not use the known gold standard (*i.e.*, pipette values) to determine this relationship. We also performed conventional regression analysis using the gold standard (*i.e.*, pipette values) for comparison. In Table 1 we summarize these results. Note that there are differences between the slopes, intercepts, and noise terms obtained from these two methods. However, the ordering of the slopes and noise terms is the same between the two methods. Regression analysis performs better than RWT because it has access to the x-coordinates (*i.e.*, the gold standard).

In Fig. 4 we plot the volume estimates obtained using the three aforementioned techniques versus the gold standard. Also shown in Fig. 4 are the lines representing the results of RWT. The results shown in Fig. 4 are

Table 1. Estimates of the linear model parameters using regression analysis with and without truth. Note that the CTI estimates were obtained using manual thresholding and magnification correction while the CTII estimates were obtained using a fixed threshold and no magnification correction.

Estimation Techniques	SPECT	CTI	CTII
	\hat{a}_{SI}	\hat{a}_{CTI}	\hat{a}_{CTII}
Estimates from regression analysis	0.9387	1.0462	1.6135
Estimates using no-gold-standard analysis	0.8091	0.9032	1.3947
	\hat{b}_{SI}	\hat{b}_{CTI}	\hat{b}_{CTII}
Estimates from regression analysis	0.0110	-0.0684	-0.0210
Estimates using no-gold-standard analysis	0.0253	0.0026	0.3572
	$\hat{\sigma}_{SI}$	$\hat{\sigma}_{CTI}$	$\hat{\sigma}_{CTII}$
Estimates from regression analysis	0.0351	0.0428	0.0646
Estimates using no-gold-standard analysis	0.0478	0.0512	0.1003

somewhat misleading given that we use the gold standard in the plots; an advantage RWT does not have. This explains the noticeable imperfections in the plots.

RWT also returns estimates of the parameters defining the underlying distribution of the gold standard. Because we generated the gold standard from a known distribution, we can, again, evaluate the performance of RWT. Figure 5 contains plots of the true and estimated densities along with a histogram of the data used in the experiment.

5. DISCUSSION AND CONCLUSION

We have further evaluated our method (RWT) for comparing estimation methods without the use of a gold standard by performing volume estimation using a phantom and multiple imaging systems. We have found that our method does, in fact, allow for the comparison of estimation techniques without the use of a gold standard. Specifically, the estimates of the linear model parameters obtained using RWT are closely correlated with those obtained through standard regression analysis using the x-axis. The errors observed in our estimates of the linear model parameters are consistent with the results of simulation studies presented in earlier works [1, 2].

The estimation tasks (SPECT, CTI, and CTII) we employed for volume estimation were not particularly noisy, as can be seen in Fig. 4. However, the slope of CTII is substantially greater than one due to magnification in our CT system. RWT accurately determined this increased slope (Fig. 4(c)).

The results of previous simulation studies using RWT indicated significant improvement with increased sample size. In future work, we intend to increase the sample size of our validation experiment as well as adding a noisy estimation technique. The lack of noise present in the three volume estimation techniques used in our experiment led to differences between the estimates of $\hat{\sigma}_m/\hat{a}_m$ for each technique that were not statistically significant. This result implies that the three volume estimation techniques we used performed equally well. We also intend to compare the estimation capabilities of different reconstruction and segmentation algorithms.

ACKNOWLEDGMENTS

This work was supported by NSF grant 9977116 and NIH grants P41 RR14304, KO1 CA87017-01, and RO1 CA 52643. This research of Todd E. Peterson, Ph.D. is supported in part by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

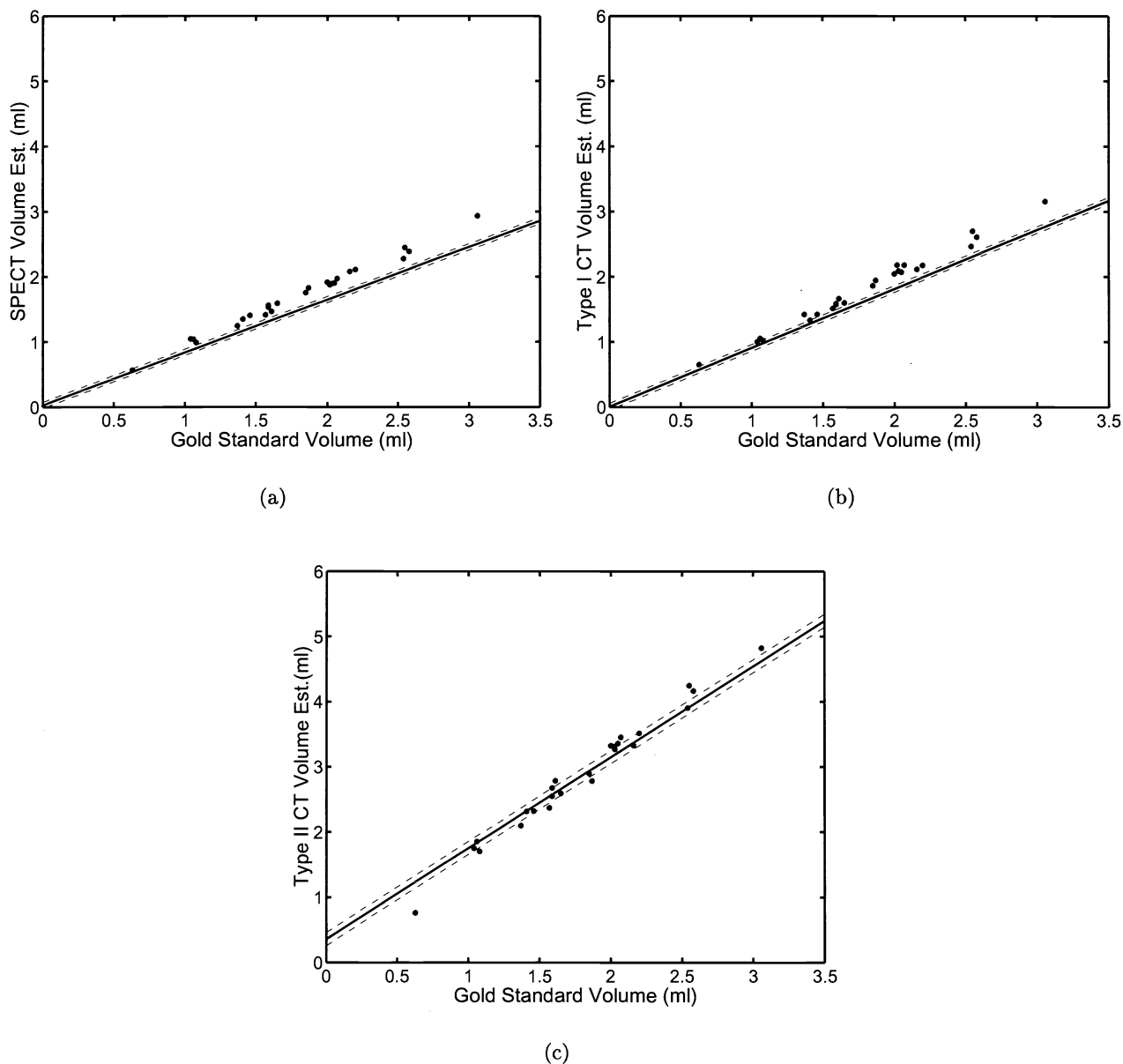


Figure 4. The results of a phantom study for estimating volumes from three estimation techniques. Twenty-five different volumes were imaged on two different modalities (SPECT, CT). In each graph we have plotted the true volume against the estimates from three different estimation techniques ((a)SPECT, (b)CTI, (c)CTII). The solid line was generated using the estimated linear model parameters for each estimation technique. The dashed lines denote the estimated standard deviations for each estimation technique. Values for these parameters are shown in Table 1. Note that although we have plotted the true volumes on the x-axis of each graph, this information was not used in computing the linear model parameters

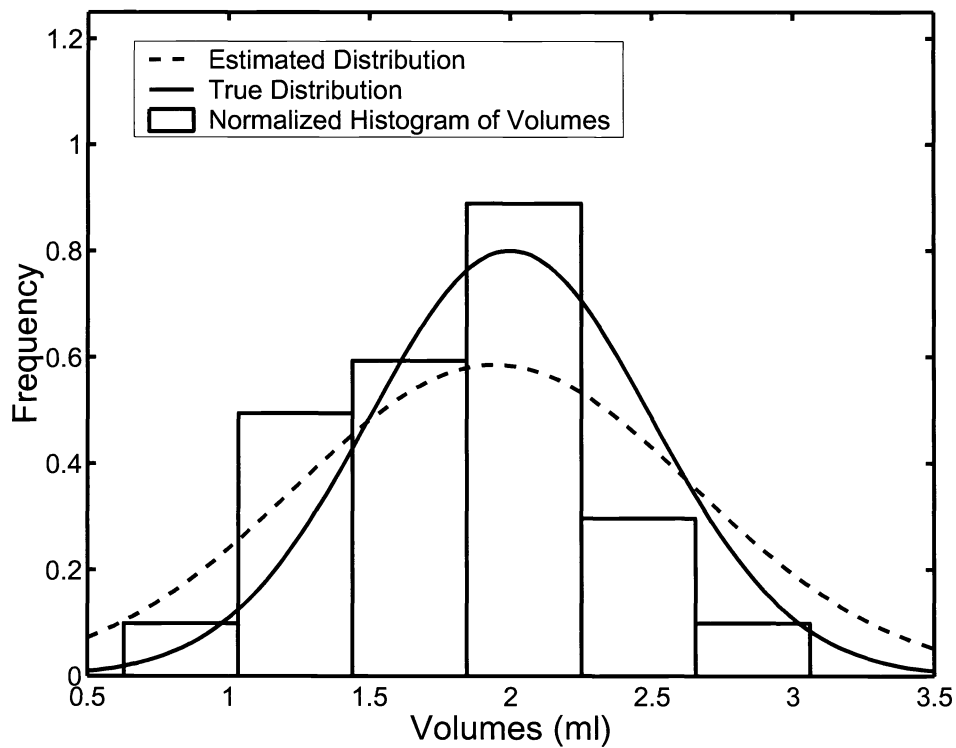


Figure 5. A comparison of the normalized histograms for the underlying volumes with the parameters returned by RWT estimating the mean and variance of the underlying gold-standard distribution. The true volumes were sampled from a truncated-normal distribution with a mean of 2ml and standard deviation of 0.5ml. The 25 volumes themselves had a sample mean of 1.8016ml and sample standard deviation 0.5607ml. Our no-gold-standard analysis predicted a mean of 1.9444ml and a standard deviation of 0.7057ml.

REFERENCES

1. J. W. Hoppin, M. A. Kupinski, G. A. Kastis, E. Clarkson, and H. H. Barrett, "Objective comparison of quantitative imaging modalities without the use of a gold standard," *IEEE Transactions on Medical Imaging* **21**, pp. 441–449, 2002.
2. M. A. Kupinski, J. W. Hoppin, E. Clarkson, H. H. Barrett, and G. A. Kastis, "Estimation in medical imaging without a gold standard," *Academic Radiology* **9**, pp. 290–297, March 2002.
3. R. M. Henkelman, I. Kay, and M. J. Bronskill, "Receiver operator characteristic (ROC) analysis without truth," *Medical Decision Making* **10**, pp. 24–29, 1990.
4. S. V. Beiden, G. Campbell, K. L. Meier, and R. F. Wagner, "On the problem of ROC analysis without truth: The em algorithm and the information matrix," in *Medical Imaging 2000: Image Perception and Performance*, **3981**, pp. 126–134, SPIE, 2000.
5. Y. Qu, M. Tan, and M. H. Kutner, "Random effects models in latent class analysis for evaluating accuracy of diagnostic tests," *Biometrics* **52**, pp. 797–810, September 1996.
6. P. S. Albert, L. M. McShane, and J. H. Shih, "Latent class modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors," *Biometrics* **57**, pp. 610–619, June 2001.
7. H. Al-Hallaq, J. N. River, M. Zamora, H. Oikawa, and G. S. Karczmar, "Correlation of magnetic resonance and oxygen microelectrode measurements of carbogen-induced changes in tumor oxygenation," *International Journal of Radiation Oncology, Biology, and Physics* **41**(1), pp. 151–159, 1998.
8. P. O. Alderson, D. F. Adams, B. J. McNeil, R. Sanders, S. S. Siegelman, H. J. Finberg, S. J. Hessel, and H. L. Adams, "Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: A prospective comparison," *Radiology* **149**, pp. 225–230, 1983.
9. M. Abe, Y. Kazatani, H. Fukuda, H. Tatsuno, H. Habara, and H. Shinbata, "Left ventricular volumes, ejection fraction, and regional wall motion calculated with gated technetium-99m tetrofosmin spect in reperfused acute myocardial infarction at super-acute phase: Comparison with left ventriculography," *Journal of Nuclear Cardiology* **7**, pp. 569–574, November/December 2000.
10. N. G. Bellenger, M. I. Burgess, S. G. Ray, A. Lahiri, A. J. S. Coats, J. G. F. Cleland, and D. J. Pennell, "Comparison of left ventricular ejection fraction and volumes in heart failure by echocardiography, radionuclide ventriculography and cardiovascular magnetic resonance," *European Heart Journal* **21**, pp. 1387–1396, August 2000.
11. E. Cwajg, J. Cwajg, Z.-X. He, W. S. Hwang, F. Keng, S. F. Nagueh, and M. S. Verani, "Gated myocardial perfusion tomography for the assessment of left ventricular function and volumes: Comparison with echocardiography," *Journal of Nuclear Medicine* **40**(11), pp. 1857–1865, 1999.
12. T. L. Faber, J. Vansant, R. I. Pettigrew, J. R. Galt, M. Blais, G. Chatzimavroudis, C. D. Cooke, R. D. Folks, S. M. Waldrop, E. Guartler-Krawczynska, M. D. Wittry, and E. V. Garcia, "Evaluation of left ventricular endocardial volumes and ejection fractions computed from gated perfusion spect with magnetic resonance imaging: Comparison of two methods," *Journal of Nuclear Cardiology* **8**, pp. 645–651, November/December 2001.
13. Z. He, E. Cwajg, J. S. Presian, J. J. Mahmarian, and M. S. Verani, "Accuracy of left ventricular ejection fraction determined by gated myocardial perfusion spect with tl-201 and tc-99m sestamibi: Comparison with first-pass radionuclide angiography," *Journal of Nuclear Cardiology* **6**, pp. 412–417, July/August 1999.
14. D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *The Statistician* **32**, pp. 307–313, 1983.
15. J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet* **i**, pp. 307–310, 1986.
16. G. A. Kastis, L. R. Furenlid, D. W. Wilson, T. E. Peterson, H. B. Barber, and H. H. Barrett, "Compact CT/SPECT small-animal imaging system," in *Nuclear Science Symposium and Medical Imaging Conference*, IEEE, November 2002.
17. G. A. Kastis, H. B. Barber, H. H. Barrett, S. J. Balzer, D. Lu, D. G. Marks, G. Stevenson, J. M. Woolfenden, M. Appleby, and J. Tueller, "Gamma-ray imaging using a cdznte pixel array and a high-resolution, parallel-hole collimator," *IEEE Transactions on Nuclear Science* **47**, pp. 1923–1927, December 2000.