Terahertz pulsed imaging of freshly excised human colonic tissues.

Caroline B. Reid[1], Anthony Fitzgerald[2], George Reese[3], Robert Goldin[4], Paris Tekkis[3], P. S. O'Kelly[5], Emma Pickwell-MacPherson[6], Adam P. Gibson[1], Vincent P. Wallace[2]

[1] *Department of Medical Physics and Bioengineering, University College London, London, WC1E 6BT*

[2]*School of Physics, The University of Western Australia, Crawley 6009, Australia and TeraView Ltd Platinum Building, St John's Innovation Park, Cowley Road, Cambridge, UK, CB4 0WS*

[3]*Division of Surgery, Chelsea and Westminster Campus, Imperial College London, London, UK*

[4]*Centre for Pathology, Imperial College London, St. Mary's Campus, London, UK*

[5]*TeraView Ltd, Platinum Building St John's Innovation Park, Cowley Road, Cambridge, UK, CB4 0WS*

[6]*Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*

Abstract

We present the results from a feasibility study which measures properties in the terahertz frequency range of excised cancerous, dysplastic and healthy colonic tissues from 30 patients. We compare their absorption and refractive index spectra to identify trends which may enable different tissue types to be distinguished. In addition, we present statistical models based on variations between up to 17 parameters calculated from the reflected time and frequency domain signals of all the measured tissues. These models produce a sensitivity of 82% and a specificity of 77% in distinguishing between healthy and all diseased tissues and a sensitivity of 89% and a specificity of 71% in distinguishing between dysplastic and healthy tissues. The contrast between the tissue types was supported by histological staining studies which showed an increased vascularity in regions of increased terahertz absorption.

Keywords: cancer, terahertz, endoscopy, colonoscopy, parametric analysis

Introduction

Colorectal cancer is the third most common cancer across the UK population with approximately 35,000 new cases per year; it is the second most common cause of cancer-related death and it is the third most common cancer across the world (1). The disease is slow to develop, for example, the risk of invasive cancer developing within an adenoma is approximately 2% per year. Therefore, early diagnosis and removal of these lesions is an effective method of reducing cancer risk. Survival depends on the stage of disease at time of presentation, where survival rates are improved if dysplastic lesions can be detected and treated early. This is the aim of colonic screening on otherwise asymptomatic patients.

The terahertz (THz = 10^{12}Hz) frequency range, covering wavelengths of 0.3mm to 3mm, is a non-ionising radiation which demonstrates a combination of good chemical specificity and depth resolution (2). The potential of THz imaging as an endoscopic tool for the detection of colonic tissue pathologies is encouraged by positive results in discriminating healthy tissues from diseased tissues in THz imaging of skin (3) (4) (5) and breast (6) (7) tissue and in advances in miniaturisation of devices (8) (9) (10). THz imaging of both skin and breast cancers have shown that the contrast is sufficient to identify tumour margins when compared with histology. Positron emission tomography (PET), magnetic resonance imaging (MRI) and near infra-red (NIR) studies have all offered evidence that tumours have an increased water content in comparison to healthy tissues (11) (12) (13) (14). Water, a polar molecule, is highly absorbing in the THz region and is thus a likely source of image contrast. Studies of dehydrated tissue samples, however, have presented evidence of contrast in THz imaging of these samples, suggesting that the contrast observed may be due to more complex mechanisms than simply water absorption, such as protein density, cellular structures or other molecular changes (15) (16). One recent study has shown that, for liver cirrhosis, 50-66% of the contrast is due to structural changes (17) (18).

The work presented in this paper investigates the ability of THz pulsed imaging to map regions of diseased tissues from healthy tissues in freshly excised human colon tissue and to determine areas of healthy from dysplastic tissues. Dysplastic tissues show abnormal development which may, in time, develop into cancerous tissue. We present the results from an imaging study where cancerous, dysplastic and non-neoplastic colonic samples were imaged from 25 patients. Twelve of these cases were further analysed through histological staining methods to explore likely contrast mechanisms in the THz images. The THz properties of excised normal and diseased tissues are presented from which parameters of the tissues are derived. We then show that the THz parameters can be used to categorise the samples into normal and diseased tissues.

Materials and methods

Sample Preparation

The tissue samples investigated in this study were from 30 patients undergoing colonic resection surgery at St. Mary's Hospital, London, UK. Ethical approval for the study was granted by the Local Research Ethics Committee and all patients had given their informed, signed consent for the excised tissues to be used for research. The samples included cancer and dysplasia and normal tissue samples were also imaged for comparison. The majority of these samples were from the lower gastrointestinal (GI) tract (the colon), with 4 cases examining upper GI cancer.

Tissue was retrieved immediately following excision from the patient and taken to the pathology department. Pathological assessment was made of the excised tissue to identify sites of diseased and healthy tissue, from which tissues could be harvested for imaging. The number of tissue samples retrieved was decided by a consultant histopathologist and was dependent on a number of factors, where the aim was to obtain tissue with minimal disruption to standard pathology procedure. Typically, two samples of tissue, one sample of healthy tissue and one sample of diseased tissue (later confirmed by histology), were imaged from each excised colon. Healthy tissue samples were harvested well away from the diseased areas.

Imaging System

The reflection imaging system used in this study was the stand-alone portable THz system TPIimaga1000 (TeraView Ltd, Cambridge, UK) a full description of the system operation is given elsewhere (19). The THz pulses are focused onto the top of a 2mm thick z-cut quartz window where the sample under investigation is placed. The reflected beam is focused onto a photoconductive device for detection. To acquire the reflected data over an area, the entire THz optics, and hence the beam, is raster-scanned in the x-y plane; an entire THz waveform is acquired at each x-y point. All measurements were made at room temperature.

Solid sample cell.

Figure 1 Shows a photograph of the tissue holder in place on the THz imaging system.

A tissue holder, shown in fig. 1, was designed to control the orientation of the tissue samples throughout the tissue analysis process, from initial imaging through tissue fixing and histological examination. The tissue holder can easily be removed from the imaging system to allow for the samples to be formalin fixed in the same position as they were imaged. A plastic lattice covering was placed over the sample and gently fixed into place with four screws to ensure the samples are securely positioned. The lattice covering is fixed gently to not overly compress the tissues which would affect their optical properties. Following sample imaging the whole tissue holder could be removed from the scanner and placed in formalin where the lattice covering allows for even fixing of the tissue sample. This helps reduce the effect of uneven tissue shrinkage which occurs during the fixing process and would impede accurate registration between histology and THz results. Gold markers on the quartz plate, which can be seen both in THz and photographic images, aid registration of the samples.

Data Acquisition and histological assessment

A reference measurement, used to normalise the measured data, was created by focussing the beam onto the top surface of the quartz plate. Both quartz/air and quartz/water references were

used for these measurements. Quartz/air measurements (referred to here as 'air references') are typically used as the reference for tissue measurements. In addition, quartz/water measurements (referred to here as 'water references') were used as the optical properties of water and tissue are similar and, so, in using the water reference to normalise the measured pulses a more sensitive measure of the change in sample properties may be achieved. Background measurements were taken for all tissue measurements and the data corrected for by subtracting the background waveform from the reference and sample waveforms (20). The acquired reflection mode waveform for each pixel represented the impulse function of the sample convolved with the reference waveform, which includes instrumental and environmental aspects such as oscillations in the waveform shape due to atmospheric water. These can be removed by subtracting the background and then deconvolving the raw waveform with the reference waveform to extract the impulse function. The Fourier transform of this time domain signal gives a frequency domain spectrum (21). A more detailed description of these mathematical methods may be found in the supporting literature (22).

Single samples of healthy tissues were placed next to individual samples of diseased tissues from the same patient on the imaging window for each imaging procedure. The time from tissue excision to imaging was typically around 30 minutes and any tissue awaiting imaging was kept in an air-tight container. Two different approaches to positioning the tissue were used; initially, (for the first eleven samples) the tissues were imaged by placing them directly onto a fixed quartz window. The tissue was photographed whilst in position on window using a camera system mounted inside the imaging system. The photograph, therefore, correlated to the acquired THz image of the tissue. The orientation of the tissue on the scan window was registered to the photographic image to enable correlation between regions of contrast in the THz image and features seen in the photographic image. The normal and tumour tissues were marked on the photo (later confirmed by histology). Once data collection had been completed, the tissue was pinned to a cork block to maintain the tissue size and shape, photographed and placed in a labelled formalin pot for fixation for 24hrs. Slight changes in tissue orientation, however, could occur during the pinning to the cork board which made image registration sometimes difficult and introduced errors. For the remaining 24 samples we improved the registration by imaging the tissues using the tissue holding device previously described (Fig. 1) which facilitated a simple tissue registration of the THz images with the subsequent histopathologic examination. Following imaging the samples were placed in a labelled formalin pot while still positioned in the tissue holding device for fixation for 24hrs.

A histopathological examination was performed to identify the size and shape of tissue pathology regions on the THz images. To carry this out on the fixed tissue, the whole sample was sequentially blocked. The sites of the histological blocks were marked on the digital image and sections were cut for histological examination to identify areas of normal mucosa, dysplasia or invasive carcinoma. The boundaries between the different pathologies were noted on the digital images.

Areas of normal, malignant and dysplastic tissues were identified from pathology results and highlighted on the pathology images. These images were then aligned with the THz images by image registration following which regions of interest (ROI) corresponding to the normal, tumorous and dysplastic regions could be marked out on the THz images. The refractive index and absorption coefficient were extracted from the raw reflection data at each individual pixel following the equations given in reference (23). These values were then averaged across the ROI to achieve a

single value describing the averaged optical properties of the reflected THz waveforms from that particular tissue pathology.

Statistical analysis

As a THz pulse is reflected from a material it will experience delay, attenuation and broadening, with respect to the reference pulse, as the various component frequencies are phase shifted, absorbed, reflected, and scattered by the interrogated material. This results in visually identifiable difference in the reflected pulse. To extract information from the measured THz pulses and create images, parameters (from both the time and the frequency domains) are identified which relate to the physical processes responsible for causing delay, attenuation and broadening of the input THz pulses.

The aim of the statistical analysis was to construct a model for the classification of normal, dysplasia or cancer using a series of THz parameters. A number of parameters for both air and water referenced data were identified and are given in table 1. The parameters identified for use in this study include those that have been used previously to distinguish tumours and healthy tissue for a number of cancers (24) (6), and it should be noted that not all of these parameters are independent of each other, some are correlated in some degree with the others. The time parameters were selected, as in previous studies (24), to provide the best contrast between the diseased and normal tissues. A full description of the selected parameters is given in Appendix A.

Table 1 Parameters for both air and water referenced data for the classification of normal, dysplastic and cancerous colonic tissues.

	Air reference parameter	Water reference parameter
1	Emin: Minimum value of THz pulse	Emax: Maximum value of THz pulse
2	-Emin/Emax	A(t): Amplitude of pulse at time t = 0.26 ps prior to Emax time
3	A(t)/Emin: Normalised amplitude at t = 0.33 ps prior to Emin time.	A(t): Amplitude of pulse at time t = 0.26 ps after Emax time
4	A(t)/Emin: Normalised amplitude at t = 2 ps after Emin time	Integral in THz pulse from 0.26 ps before Emax time to 0.26 ps after
5	Integral of values in THz pulse from 0.33 ps before Emin time to 2 ps after.	P(f): Power in spectrum at f = 0.6 THz
6	P(f): Power in spectrum at f = 0.6 THz	Integral of power spectrum over freq range 0.44 - 0.60 THz
7	Integral of power spectrum over freq range 0.44 - 0.60 THz.	Log of Integrated power
8	Full width half maximum (FWHM)	Full width half maximum (FWHM)
9	Emax - Emin	

Parametric images were generated for each parameter for each waveform within the previously highlighted tissue ROI's, which were then averaged to generate an averaged value for the ROIs. These parametric values were then analysed using STATA and SPSS, two general-purpose statistical software packages. The averaged waveforms were grouped into benign (normal pathology) and neoplastic pathology (cancer and dysplasia). As the aim of colonic screening is to identify regions of the pre-cancerous dysplastic tissues, it is useful to assess how accurately these regions can be identified using THz imaging. Firstly a multinomial logistic regression analysis was used to compare the parameters from all ROIs. Secondly binary logistic regression analysis was used to compare parameters only from regions identified as healthy and dysplastic.

Slide staining

To investigate other potential sources of contrast in the samples, staining of the slides was undertaken to identify a physical or chemical property of the tissue which could then be correlated to differences in THz measurements between pathologies. Twelve samples were identified for staining using histochemical and imunohistochemical stains. A list of the stains used is given in table 2. Masson's trichrome and sirius red are two stains used to identify the collagen content of the samples. Diastase resistant periodic acid Schiff (dPAS) was used to identify mucin which is a characteristic feature of adenocarcinomas, such a colonic cancers. Platelet Endothelial Cell Adhesion Molecule (CD31) was used to locate the endothelium of small vessels and identify regions of neoangiogenesis which is a critical step in the progression of neoplasia (25). CAM 5.2 is a broad spectrum anticytokeratin (specifically cytokeratins 8, 18 and 19) which stains most epithelial neoplasms and was used to identify carcinoma cells (26). Immunohistochemical staining for D240 was used to identify lymphangiogenisis.

The slides were manually examined using a microscope and regions of pathology were identified from histology.

Table 2 Stains used in colon histology analysis

Stain	Identifies
MT	Collagen
SR	Collagen
DPAS	Mucin
CD31	Endothelial marker
CAM 5.2	Epithelial cells.
D240	Lymphatic marker

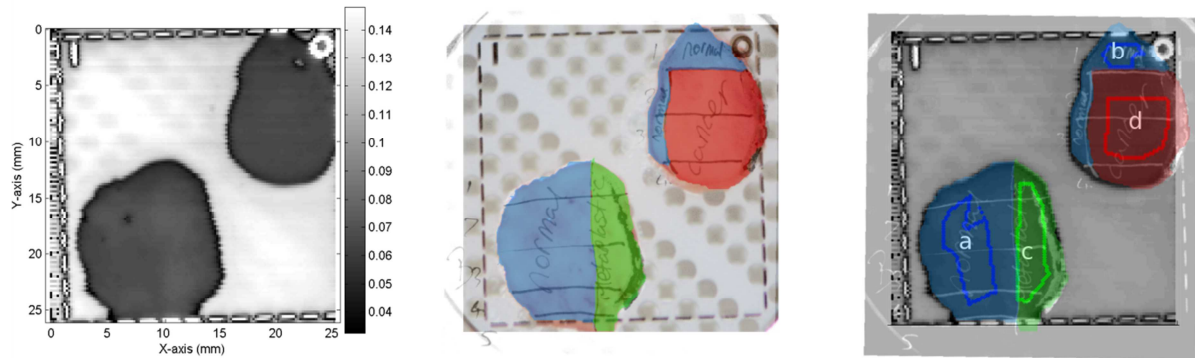# Results

Imaging and histology results

Figure 2 Illustration of the imaging procedures used in the colon tissue study. a) is an example THz image of tissue containing healthy regions, dysplasia and and cancerous tissue. In this particular case the THz image generated from the E_{min} signal is shown. b) shows the histology results (drawn onto a photographic image of the tissue samples). c) The histology results are overlaid on the THz image from which regions of interest can be selected. Impulse functions for the respective tissue types are extracted from the regions of interest. In this example, region A and B are normal tissue, C is dysplastic tissue and D is cancerous tissue. Figures b) and c) have been shaded to indicate each region; blue = normal tissue, green = dysplastic tissue and red = cancerous tissue. A description of the method of selection of these regions is given in the Data Acquisition and histological assessment section .
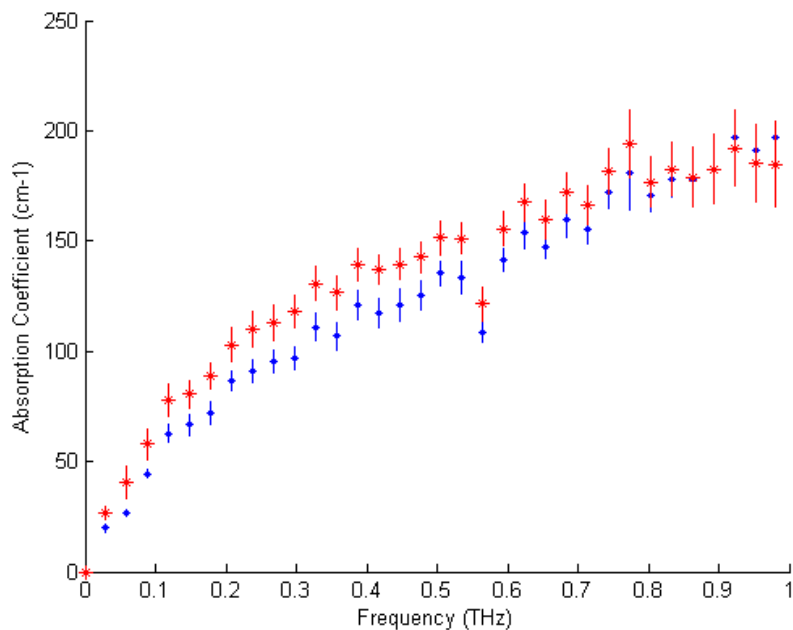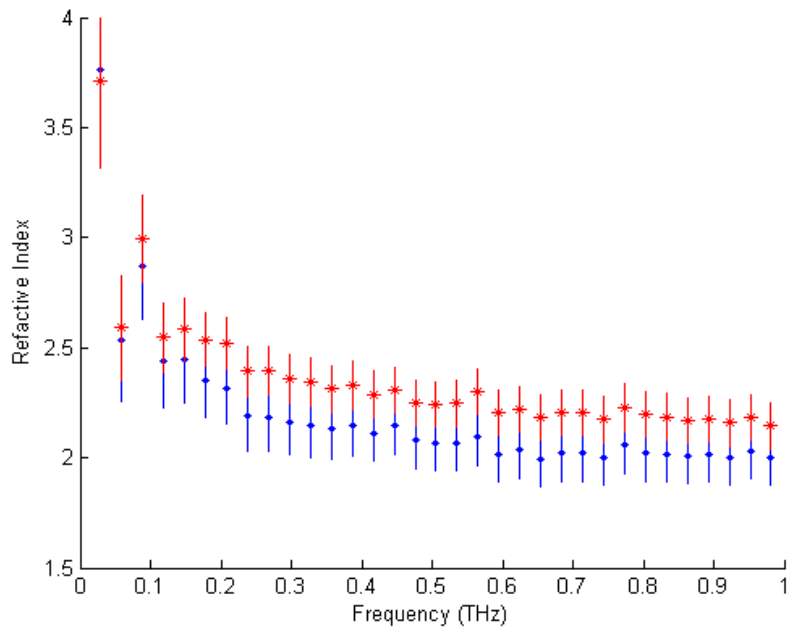
Figure 3a
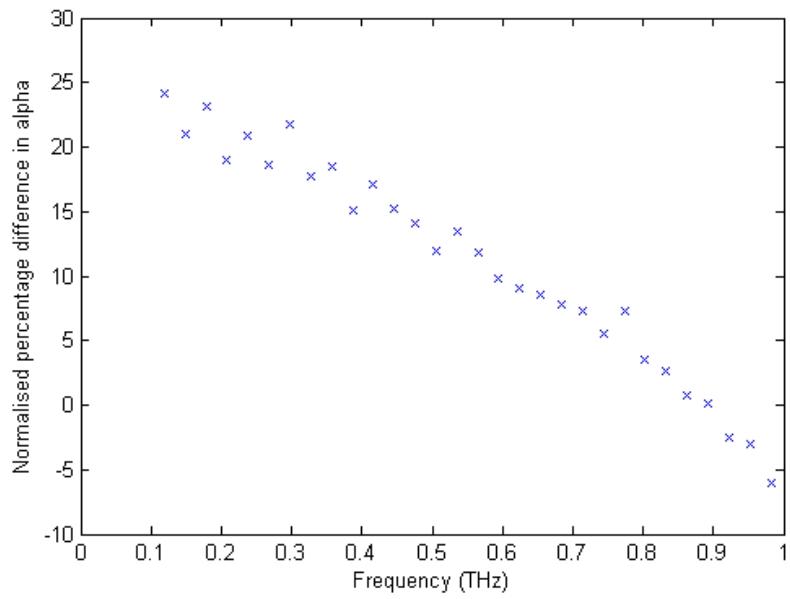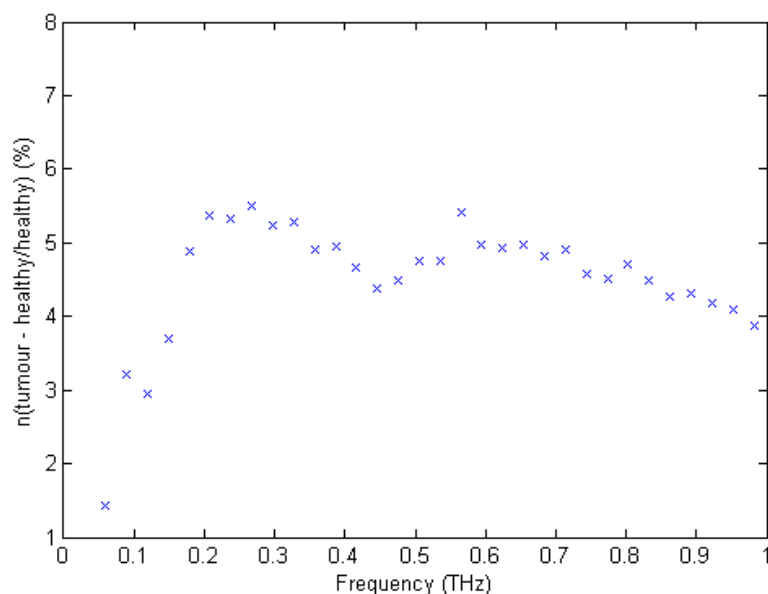

Figure 3b

Figure 3c

Figure 3d

Figure 3 The mean optical properties, with standard error for combined healthy tissues (the full circle, coloured blue) for all patients compared to the properties for tissue with tumour (the star, coloured red). a) Shows the mean absorption coefficient together with the standard error. b) Shows the refractive index. c) Shows the percentage difference of the mean absorption coefficient for normal and tumour tissue, for all patients. d) Illustrates the percentage difference of refractive index for normal and tumour tissue for all patient tissues. Results are shown up to 1 THz as above this the influence of noise from phase effects becomes more dominant.

An image is shown in fig. 2, from which it can be seen that the cancerous and dysplastic tissue regions show contrast relative to the normal tissue, indicating differences in the reflected waveforms.

Absorption coefficients and refractive indices were calculated from the reflected waveforms for cancerous and healthy tissues for those samples held in the tissue sample holder. As the accuracy of the measurements may be affected by difficulties in determining the true optical properties of a material in reflection mode, only samples measured in the sample-holder were used in this particular analysis as these sources of error are minimised in comparison to those samples not measured in the sample-holder. The optical properties determined from reflection mode measurements of the THz imaging system are derived from the phase components of the imaginary data. The phase is highly sensitive to sample position meaning any difference between the reference and the raw data measurement will result in errors in the measured optical properties. Furthermore, the reference signals were made at the centre point of the imaging window and, as the reference was applied across the whole image regardless of location, this could lead to errors in the measured phase at the non-central points in the image (27). Other potential sources of error in the measurements could be the imaging window not being entirely flat causing the phase to change across the window, the sample not being in close contact across the scanned surface, pressure on the sample creating a bend in the supporting metallic structure of the quartz window which would change the phase across the surface or pressure on the sample caused by the top surface of the tissue holding device though effort was made to minimise this effect.

In fig. 3, plots showing the patient averaged absorption coefficient and refractive index of cancer tissue are compared to that of the normal tissue, which is consistent with higher water content in

the diseased tissue. Error bars showing the standard error of the averaged values in Fig. 3 highlight the spread of measurements and the variation between samples. The difference plots shown in Figs. 3c and 3d demonstrate that at lower THz frequencies the difference between the averaged absorption coefficient and refractive index of cancer and healthy tissues are greater, which is in keeping with other THz imaging studies of excised human tissue (7). This suggests that more confidence can be given to the distinction between the derived optical properties of the tissues at the lower frequency range.

Statistical analysis

Despite the sensitivity of the calculation of the optical properties to the measurement phase, the calculation of the parameters from the reflected waveforms is not so problematical, since the highly sensitive phase part of the signal is not being used. This suggests the use of parameters in the analysis of the reflected waveforms is a more robust method than the analysis of the absolute optical properties. This analysis was performed on all tissue samples, irrespective of tissue measurement method.

For each patient 17 parameters from the reflected THz pulses in the regions of interest were averaged and compared between healthy, tumour and dysplastic tissue pathologies. Examples are shown in Fig. 4. Fig. 4a shows patient boxplots of the distributions of the air reference E_{min} values for both the healthy (shown in blue) and cancerous (shown in red) tissues. Fig. 4b shows boxplot distributions for dysplastic and healthy tissue for air reference parameter 5, the integral of values in THz pulse from 0.33 ps before Emin time to 2 ps after. It can be noted that the difference between the parameters for the healthy and diseased tissues are, though varying in the degree of observed difference due to biological spread between the samples, consistent across almost all the tissue samples.

Due to this biological spread it would be difficult to suggest that a definite distinction may be made between the tissue types when considered as a group. However, if analysis is performed for each case using the known normal tissue as a reference, it may be possible to identify differences between the tissue pathologies. This point is illustrated by the three-dimensional parametric plots in fig. 5 which show that normal and tumour tissue can be classified by considering multiple parameters.
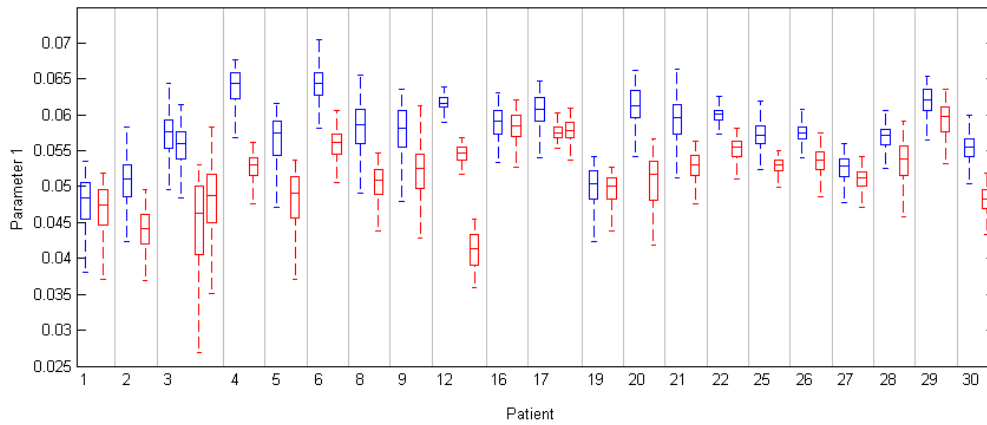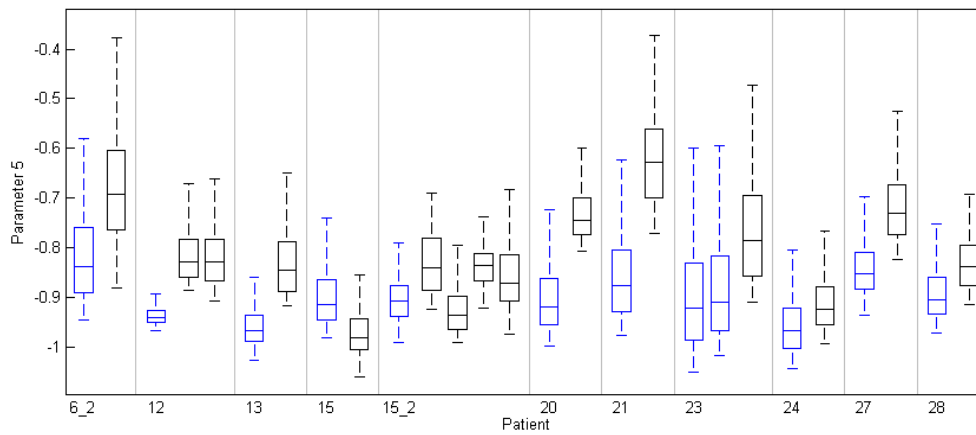
Figure 4a

Figure 4b

Figure 4 The values of parameter ranges from the regions of interest for each patient, shown as boxplots for a) the first air reference parameter, E_{min} for cancerous tissues (red) in comparison to their corresponding healthy tissue samples (blue) and b) values of the fifth air reference, P(f): Power in spectrum at f = 0.6 THz, parameter for dysplastic tissues (black) in comparison to their corresponding healthy tissue samples (blue). In the boxplots, the central marks represent the median of the parameter ranges, the edges of the boxes represent the 25th and 75th percentiles and the whiskers represent the most extreme data points of the parameter ranges.
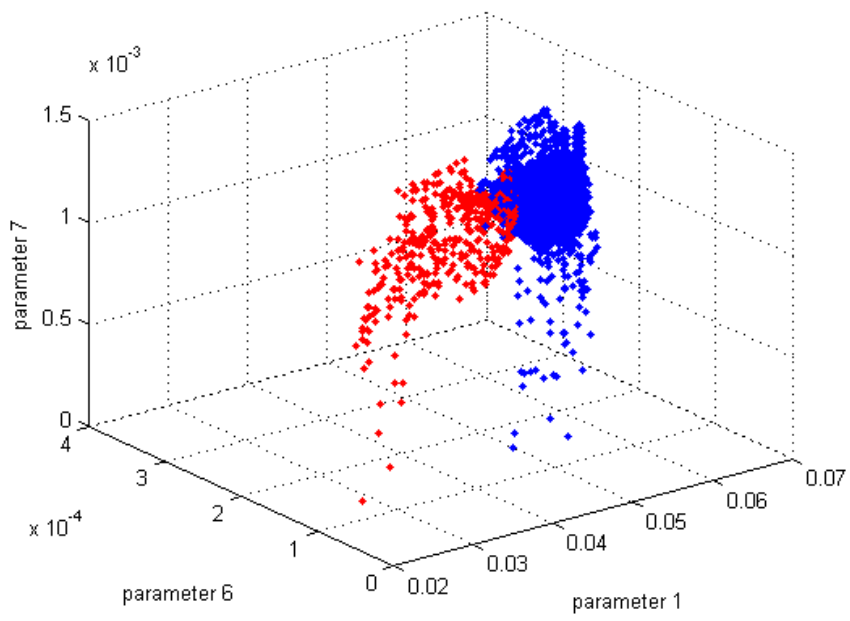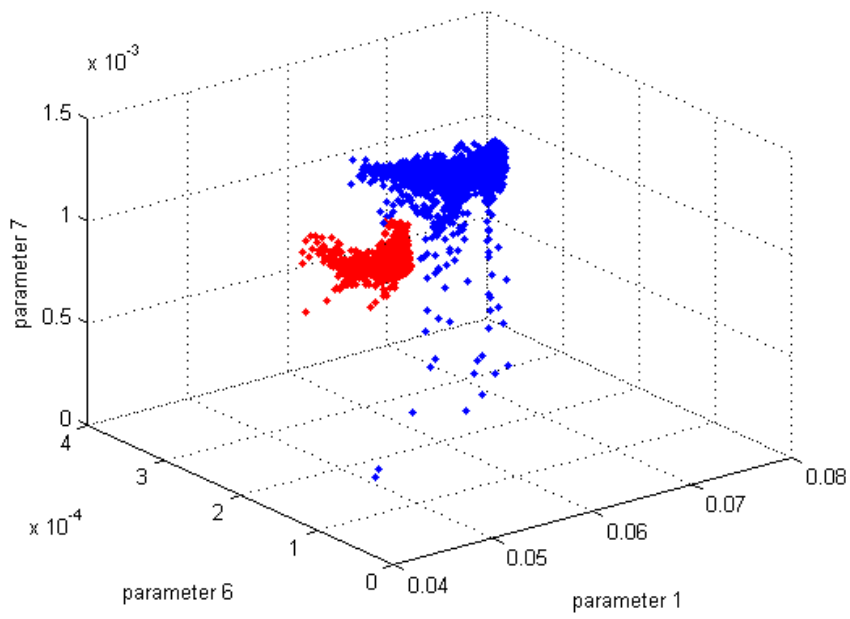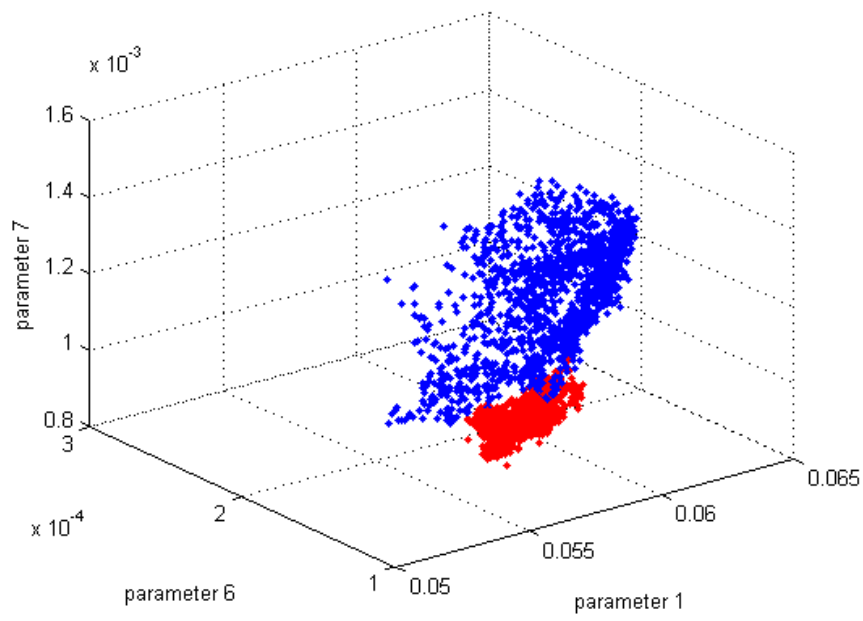
Figure 5a

Figure 5b

Figure 5c
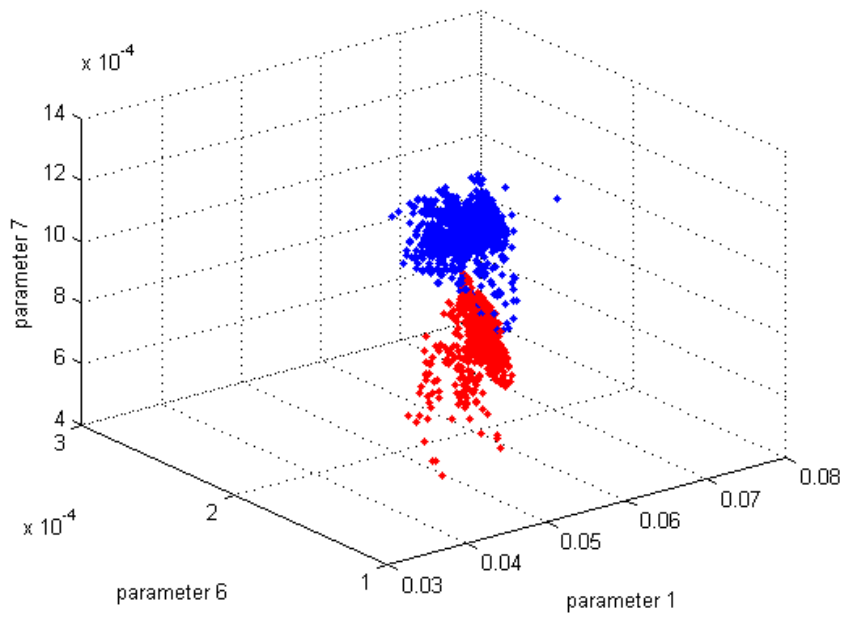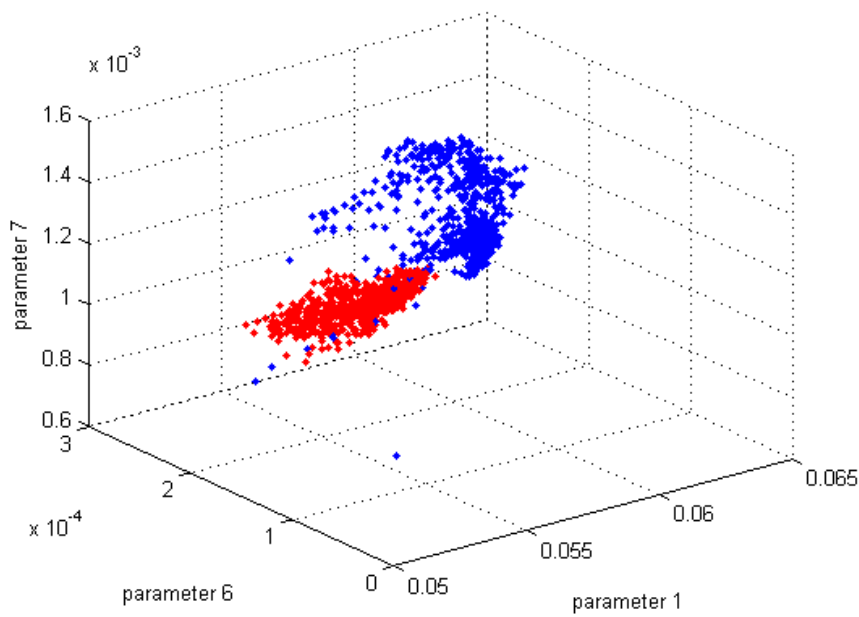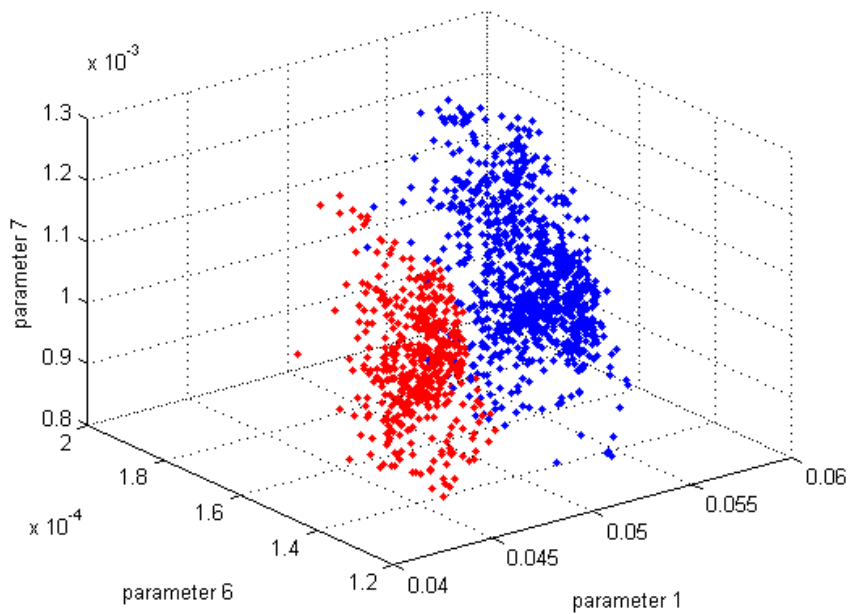
Figure 5d

Figure 5e

Figure 5f

Figure 5 Three dimensional plots illustrating how tumour (red dots) and healthy tissue (blue dots) may be classified for each patient with Terahertz imaging by using several parameters, in this case three. The patients shown are a) 3 b) 5 c) 20 d) 21 e)22 f) 30

The first statistical model used multinomial logistic regression analysis to compare parameters from healthy tissues to those from both cancerous and dysplastic tissues. Six of the 17 parameters were identified as statistically significant; however, the model gave best classification when 9 of the 17 parameters were considered. These were air references, 1, 3, 7 and 9 and water references 1, 4, 5, 6 and 8, as given in Table 1 Parameters for both air and water referenced data for the classification of normal, dysplastic and cancerous colonic tissues.. The model estimated a sensitivity of 82% and a

specificity of 77% in THz imaging in distinguishing healthy tissues from gross `diseased tissues'. The second statistical model used binary logistic regression analysis to compare parameters from healthy tissues to parameters from only dysplastic tissues. Individually most parameters showed a statistically significant difference between normal and dysplasia. The model suggested an 89% sensitivity and a 71% specificity. The model suggested that the data for the 2 groups was significantly different with a p value of 0.021.

Slide staining

Angiogenesis of the tissue samples, from normal through dysplasia to invasive carcinoma, was identified from the CD31 stain. The D240 stain, a lymphatic marker, also showed a more pronounced difference between tissue pathologies with 3.5% for the normal regions, 12% for the dysplastic regions, increasing to 14% for the cancer regions. Tumours produce a range of growth factors that directly or indirectly stimulate both blood and lymphatic vessel growth (28), which produce the increased densities lymphatic and vascular cells in the dysplastic and malignant samples. As newly formed vessels are leaky (29), positive results from both the CD31 and D240 stains may suggest an increase in liquid content of the tissues with disease progression which may be a source of image contrast. However, as the tissues were imaged following excision, we cannot be certain of the preservation of the liquid content of the lymphatic and valuator systems during imaging. An increase in vascularity and lymphatic systems also implies a change in structure in these regions. The other stains did not reveal any uniform differences between the pathology groups.

Conclusions

The work presented in this study has demonstrated the sensitivity of THz imaging to differences in colon tissue composition under controlled conditions. Our investigation into the spectroscopic properties of the *ex vivo* tissues showed an increase in both the absorption coefficient and refractive indices of the diseased tissues in comparison to healthy tissues demonstrating contrast between the tissue pathologies even when averaged over a large dataset. Through tissue staining studies, we have identified increases in lymphatic systems and vasculature of the diseased tissue in comparison to the normal mucosa, which may indicate physical differences between the tissue pathologies useful in determining underlying contrast mechanisms in THz imaging. Our statistical analysis has determined a sensitivity of 82% and a specificity of 77% when comparing parameters from the reflected THz signals of healthy and `diseased' tissues. When comparing parameters from the reflected THz signals of healthy and dysplastic tissues, a sensitivity of 89% and a specificity of 71% was returned, with a p value of 0.021. These parameters have shown consistency in their differences between the healthy and diseased tissue states which 'matching' samples from an individual patient are considered. Potential exists, therefore, for a THz based parametric technique where a normal tissue measurement is taken well away from suspected regions of disease, calibrating the system to the patient. All parametric analysis of suspected diseased regions would then be performed relative to the patient's own normal tissue area. This calibration method is similar to those employed in microwave breast probes (30)

Acknowledgements

Bibliography

1. *Cancer Research UK.* [Online] 2010. http://www.cancerresearchuk.org/.

2. *Three-Dimensional imaging of optically opaque materials using non-ionizing terahertz radiation.* **Wallace, V.P., et al.** 2008, Journal of the Optical Society of America A.

3. *Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue.* **Woodward, R.M., et al.** 2002, Physics in Medicine and Biology, Vol. 47, pp. 3853-3863.

4. *Terahertz pulsed imaging of basal cell carcinoma ex vivo and in vivo.* **Wallace, V.P., et al.** 2004, British Journal of Dermatology, Vol. 151, pp. 424-432.

5. *Terahertz Pulsed Spectroscopy of Human Basal Cell Carcinoma.* **Wallace, V.P., et al.** 10, s.l. : Journal of Applied Spectroscopy, 2006, Vol. 60, pp. 1127-1133.

6. *Terahertz pulsed imaging of human breast tumors.* **Fitzgerald, A.J., et al.** 2, 2006, Radiology, Vol. 239, pp. 533-540.

7. *Terahertz pulsed spectroscopy of freshly excised human breast cancer.* **Ashworth, P.C., et al.** 15, 2009, Optics Express, Vol. 17, pp. 12444-12454.

8. *Development of a hand-held TPI system for medical applications.* **Wallace, V.P., et al.** 2005, Microwave Symposium Digest, 2005 IEEE MTT-S International, p. 3.

9. *Metal wires for terahertz wave guiding.* **Wang, K. and Mittleman, D.M.** 7015, 2004, Nature, Vol. 432, pp. 376-379.

10. *Metal-wire terahertz time-domain spectroscopy.* **Walther, M., Freeman, M.R. and Hegmann, F.A.** 26, 2005, Applied Physics Letters, Vol. 87, pp. 261107-261107/3.

11. *Magnetic Resonance Imaging of Human Melanoma Xenografts in Vivo: Proton Spin—lattice and Spin—spin Relaxation Times Versus Fractional Tumour Water Content and Fraction of Necrotic Tumour Tissue.* **Rofstad, E.K., et al.** 3, 1994, International Journal of Radiation Biology, Vol. 65, pp. 387-401.

12. *Non-Invasive In Vivo Characterization of Breast Tumors Using Photon Migration Spectroscopy.* **Tromberg, B.J., et al.** 200, Neoplasia, Vol. 2, pp. 26-40.

13. *Measurement of the Extracellular Space in Brain Tumors Using 76Br-Bromide and PET.* **Bruehlmeier, M., et al.** 8, 2003, The Journal of Nuclear Medicine, Vol. 44, pp. 1210-1218.

14. *Effects of THz Exposure on Human Primary Keratinocyte Differentiation and Viability.* **Clothier, R.H. and Bourne, N.** 2, 2003, Journal of Biological Physics, Vol. 29, pp. 179-185.

15. *Visualisation and classification in biomedical terahertz pulsed imaging.* **Löffler, T., et al.** 2002, Physics in Medicine and Biology, Vol. 47, pp. 3847-3852.

16. *Multispectral classification techniques for Terahertz pulsed imaging: an example in histopathology.* **Berry, E., et al.** 5, 2004, Medical Engineering and Physics, Vol. 26, pp. 423-430.

17. *The Effects of Formalin Fixing on Terahertz Image Contrast.* **Sun, Y.W., Fischer, B.M. and Pickwell-MacPherson, E.** 6, s.l. : J. Biomedical Optics, 2009, Vol. 14, p. 064017.

18. *Terahertz spectroscopy of liver cirrhosis: investigating the origin of contrast.* **Sy, S., et al.** s.l. : Physics and Medicine in Biology, 2010, Vol. 55, p. 7587.

19. *Terahertz pulsed imaging and spectroscopy for biomedical and pharmaceutical applications.* **Wallace, V. P., et al.** 255, 2004, Faraday Discussions, Vol. 126.

20. *Accuracy and resolution of THz reflection spectroscopy for medical imaging.* **Reid, C.B., et al.** 16, 2010, Physics in Medicine and Biology, Vol. 55, pp. 4825–4838.

21. *Investigation of aqueous alcohol and sugar solutions with reflection terahertz time-domain spectroscopy.* **Uhd Jepsen, P., Møller, U. and Merbold, H.** 22, 2007, Optics Express, Vol. 15, pp. 14717-14737.

22. **Nuss, Martin and Orenstein, Joseph.** Terahertz time-domain spectroscopy. [editor.] George Grüner. *Millimeter and Submillimeter Wave Spectroscopy of Solids.* Berlin / Heidelberg : Springer, 1998, pp. 7-50.

23. *Tissue characterization using terahertz pulsed imaging in reflection geometry.* **Huang, S.Y., et al.** 1, 2009, Physics in Medicine and Biology, Vol. 54, pp. 149-160.

24. *Terahertz Pulse Imaging of ex vivo Basal Cell Carcinoma.* **Woodward, R.M., et al.** 1, 2003, Journal of Investigative Dermatology, Vol. 120, pp. 72-78.

25. *A multiple scale model for tumor growth.* **Alarcon, T., Byrne, H.M. and Maini, P.K.** 2, 2005, Multiscale Model Simulation, Vol. 3, pp. 440-475.

26. *The catalog of human cytokeratins: Patterns of expression in normal epithelia, tumors and cultured cells.* **Moll, R., et al.** 1, 1982, Cell, Vol. 31, pp. 11-24.

27. *A New Data Processing Approach to Reduce Variations in Terahertz Pulsed Imaging.* **Huang, S.Y., Pickwell-MacPherson, E. and Zhang, Y.T.** s.l. : Workshop on Wearable and Implantable Body Sensor Networks (BSN), 2008.

28. *Emerging mechanisms of tumour lymphangiogenesis and lymphatic metastasis.* **Cao, Y.** 9, 2005, Nat Rev Cancer, Vol. 5, pp. 735-743.

29. *PDGF and Vessel Maturation.* **Hellberg, C., Östman, A. and Heldin, C.H.** 2010, Recent Results in Cancer Research, Vol. 180, pp. 103-114.

30. *A clinical prototype for active microwave imaging of the breast.* **Meaney, P.M., et al.** 11, 2000, Microwave Theory and Techniques, IEEE Transactions on, Vol. 48, pp. 1841 - 1853.

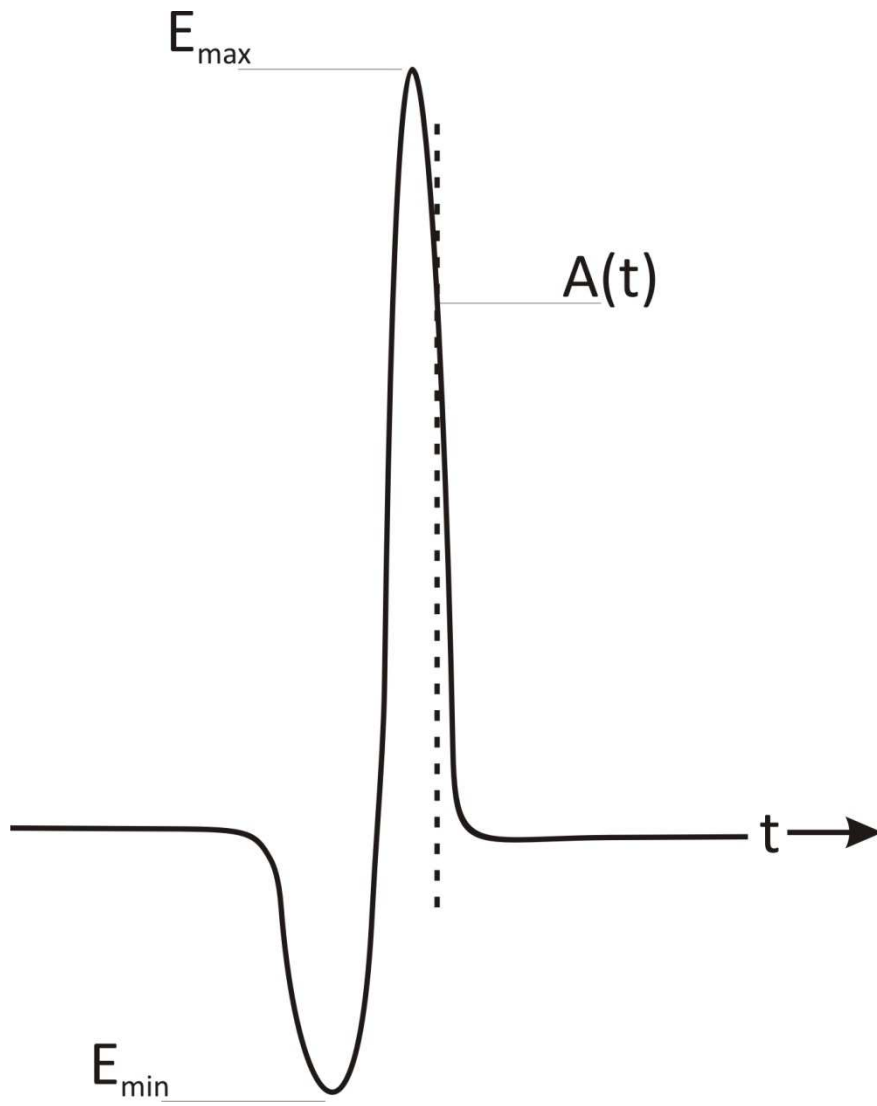# Appendix A: Description of parametric analysis
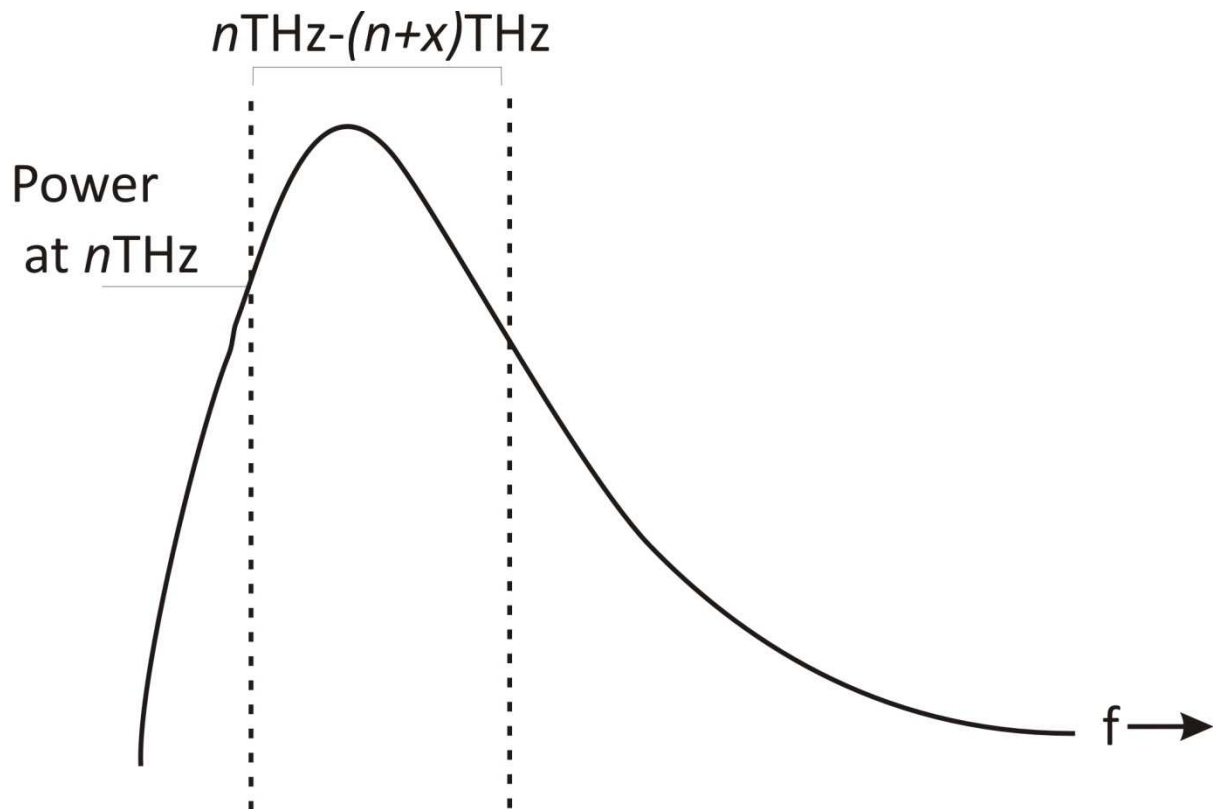
Figure 6a

Figure 6b

Figure 6 Examples of delay, attenuation, and broadening seen in the terahertz pulse and frequency spectra through differing thicknesses of safflower oil.

As a pulse is transmitted through or reflected from a material it will experience delay, attenuation and broadening, with respect to the reference pulse, as the various component frequencies are phase shifted, absorbed, reflected, and scattered by the interrogated medium. To extract information from the measured pulses and create images, parameters are identified which relate to the physical processes responsible for causing delay, attenuation and broadening of the pulses in both the time and the frequency domains. These images are referred to as 'parametric images'. The parameters used in this study, as given in table 1, consist of both time and frequency domain parameters and are illustrated in Fig 6. Figure 6a demonstrates three time domain parameters; the maximum amplitude of the time domain pulse (E_{max}), the minimum amplitude of the time domain pulse (E_{min}), and the amplitude of the time domain pulse at an arbitrary time (A(t)). From these measures, air parameters 1, 2, 3, 4, 5, 8 and 9 and water references 1, 2, 3, 4 and 8 were derived. Figure 6b demonstrates two frequency domain parameters; power at an arbitrary frequency (P at f = nTHz) and the frequency range between two arbitrary frequencies (nThz – $(n+x)$THz). From these measures, air parameters 6 and 7 and water references 5, 6 and 7 were derived.

To determine the parameters, principal components analysis (PCA) was applied to the THz difference waveforms from the tissue samples. For each of the ROIs identified by pathology, the pulses were extracted and PCA performed. The first 10 principal components (PCs) were used and applied to the pulses to derive a mean PC pulse value with standard deviation for each ROI. The steps used for the PCA and application of the PCs to the pulses to generate values useful for classification are outlined below;

1. ROIs were selected for inclusion on the basis of high SNR and on their location being away from an edge of the tissue sample.
2. Each of the selected ROIs was grouped into; normal (N), tumour (T), hyerplasia (H) or dysplaysia (D). All other pathologies were ignored.
3. PCA was performed using the matlab function *princomps*
4. Due to memory constraints, the PCA was applied to every 2nd pulse in the sample set. To confirm the PCAs were consistent with applying the method to the full dataset, the first 10 PCs were compared for using every 5th point, 3rd, point and alternate 2nd point. In each case the PCs were very similar and illustrated convergence of the PCs.
5. To calculate a value useful for classification, the first 10 PCs were matrix multiplied with the extracted pulses for each ROI. The values for each ROI were then averaged together and the standard deviation calculated. This gave a single mean value for each ROI, together with the standard deviation that could be used for classification.

Appendix B: Details of statistical analysis

Classification was achieved using both binomial logistic regression analysis and binary logistic regression analysis in SPSS 14. Data were classified into a benign (normal) group (category 0) and into a neoplastic (cancer + dysplasia) group (category 1). In addition for the multinomial logistic regression model a third category of inflamed tissues was added, for comparison purposes, referred to here as category 2. These tissues had been identified using the same pathology methods described in the body of this paper, but the results will not be discussed further. Figure 7 shows scatter plots of the standard deviations of the air and water parameters demonstrating the high correlation between the majority of the parameters.
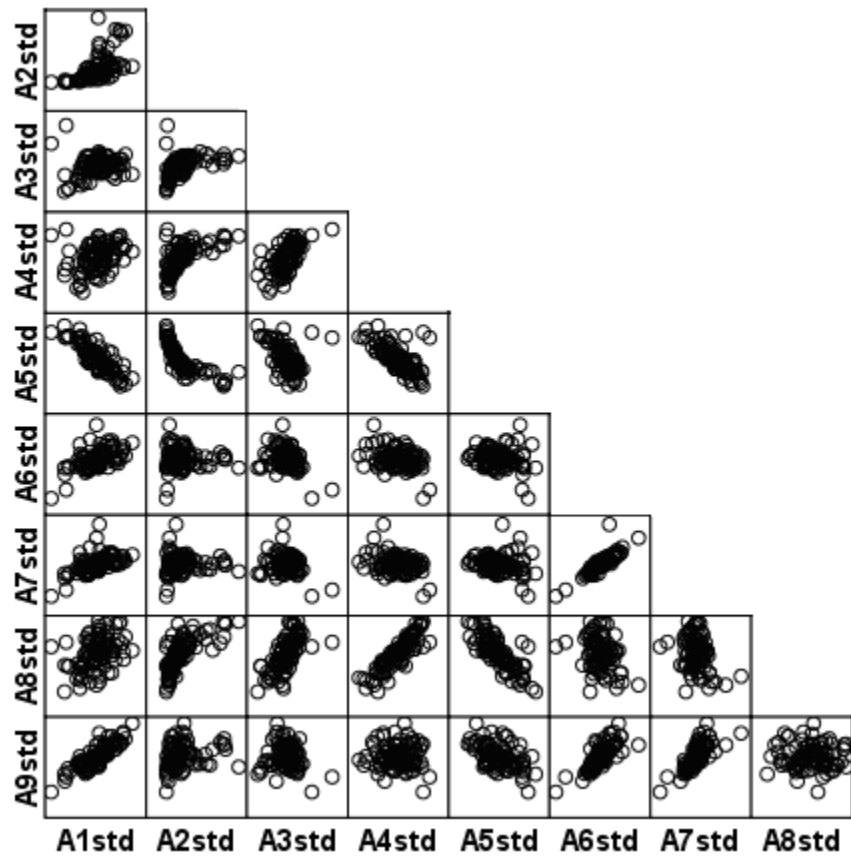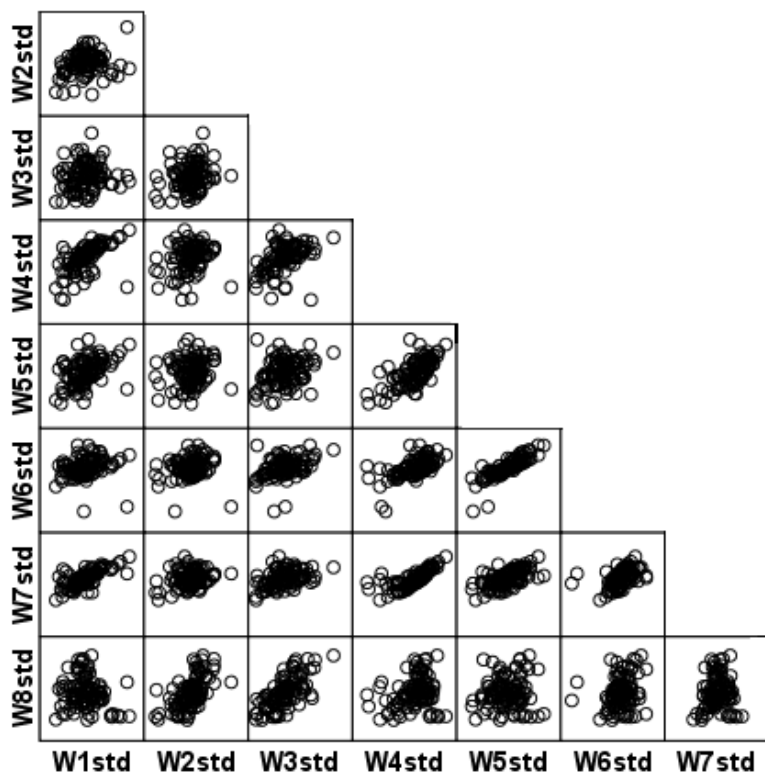
Figure 7a
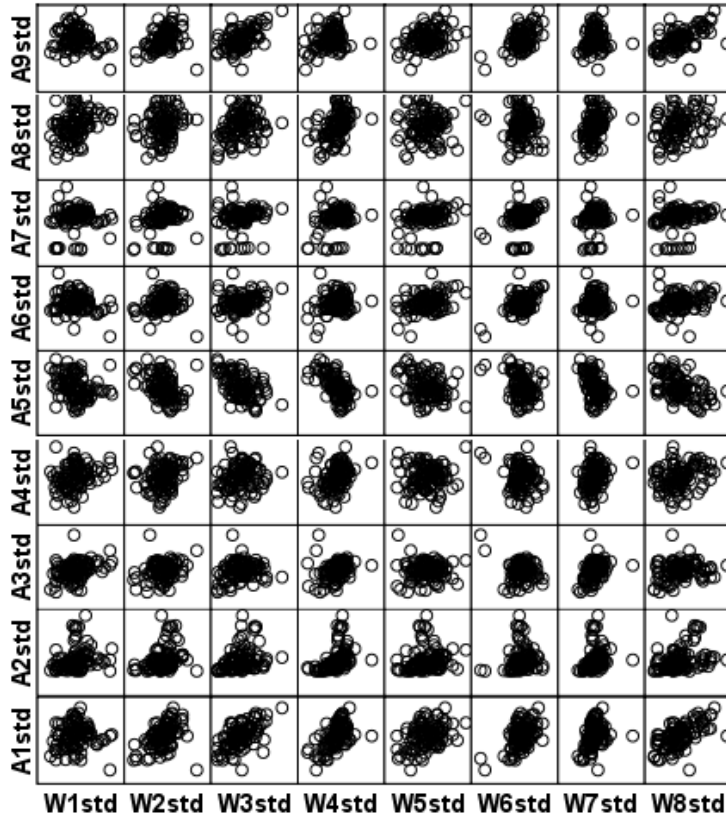
Figure 7b

Figure 7c

Figure 7 Scatter plots of the demonstrating the high correlation between the standard deviations of the air and water parameters.

1. Multinomial Logistic Regression

The method was used to take into account the nested or hierarchical structure of the data: Normal and non normal tissues within patients. In multinomial logistic models there more than two categories and a model is fitted for each pair wise comparison. For example, for categories 0, 1 and 2 there are three possible comparisons, 1 vs. 0, 2 vs. 0 and 2 vs.1, resulting in three sets of equations which are linear combinations of the air and water parameters. In calculating the multinomial logistic regression model, the predicted probabilities are calculated as follows:

$$P(Y = j) = \frac{Z_j}{1 + \sum_{j=1}^{J-1} Z_j} \qquad j : 1, \ldots, J-1$$

where J is the number of categories.

$$Z_j = \exp(\alpha_{0j} + \alpha_{1j} a_1 + \alpha_{2j} a_2 + \cdots + \alpha_{9j} a_9 + \beta_{1j} w_1 + \beta_{2j} w_2 + \cdots + \beta_{8j} w_8)$$

$$P(Y = J) = 1 - \sum_{j-1}^{J-1} P(Y = j)$$

To compute the last category (or reference category) ,

When considering the correlation within each patient the estimated effects are basically the same as the ones computed when all cases are treated as being independent but what does change is the *p*-value and the Standard Errors of the parameters estimates.

Parameter Estimates

Pathology[a]		B	Std. Error	p-value	Exp(B)
1	Intercept	1.208	.577	.052	
	A1std	2.065	1.176	.079	7.887
	A2std	-.697	.482	.148	.498
	A3std	-1.083	.373	.004	.339
	A4std	-.342	.921	.718	.711
	A7std	-1.595	1.153	.167	.203
	A8std	.480	.890	.590	1.615
	A9std	-2.193	1.111	.048	.112
	W1std	3.519	.910	.000	33.762
	W3std	1.505	.655	.022	4.505
	W4std	-6.090	2.386	.011	.002
	W5std	-2.326	.966	.016	.098
	W6std	2.558	.970	.008	12.916
	W7std	-1.065	.645	.097	.345

a. The reference category is: 0.

Table 3 Estimated effects for the comparison of normal tissue vs neoplastic tissues

Parameter Estimates

Pathology[a]		B	Std. Error	p-value	Exp(B)
2	Intercept	-4.473	2.142	.049	
	A1std	3.687	2.063	.074	39.914
	A2std	.880	.921	.340	2.410
	A3std	-3.560	1.688	.035	.028
	A4std	2.980	1.607	.064	19.695
	A7std	2.625	1.597	.100	13.811
	A8std	-2.907	1.334	.029	.055
	A9std	-2.531	1.866	.175	.080
	W1std	1.530	1.694	.366	4.620
	W3std	.300	1.681	.858	1.350
	W4std	-.684	5.652	.914	.505
	W5std	-3.413	1.554	.028	.033
	W6std	1.047	.986	.289	2.848
	W7std	2.013	1.132	.037	7.483

a. The reference category is: 0.

Table 4 Estimated effects for the comparison of normal tissue vs inflamed tissue

Parameter Estimates

Pathology[a]		B	Std. Error	p-value	Exp(B)
2	Intercept	-5.681	2.142	.012	
	A1std	1.622	2.063	.397	5.061
	A2std	1.577	.921	.122	4.838
	A3std	-2.477	1.688	.107	.084
	A4std	3.322	1.607	.025	27.714
	A7std	4.221	1.597	.026	68.090
	A8std	-3.387	1.334	.004	.034
	A9std	-.338	1.866	.870	.713
	W1std	-1.989	1.694	.178	.137
	W3std	-1.205	1.681	.489	.300
	W4std	5.406	5.652	.319	222.816
	W5std	-1.087	1.554	.540	.337
	W6std	-1.512	.986	.231	.221
	W7std	3.077	1.132	.011	21.702

a. The reference category is: 1.

Tables 3-5 shows the pair wise comparisons for normal vs neoplastic tissue, normal vs inflamed tissue and neoplastic vs inflamed tissue, respectively. These tables consist of the estimated coefficients (B), corresponding standard errors, *p-values* and the estimated odds ratios (Exp(B)). It can be seen that some variables are significant, for example, there is significant differentiation by water parameter 1 (W1). In interpreting the estimated effects, it can be seen that for a unit increase in air reference 3 (A3), the odds of category 1 vs. 0 are decreased by a factor of 0.339. In verifying how well the model classifies cases into the three categories a classification table is given in table 6.

Classification

Observed	Predicted			
	0	1	2	Percent Correct
0	27	9	0	75.0%
1	9	33	1	76.7%
2	5	0	8	61.5%
Overall Percentage	44.6%	45.7%	9.8%	73.9%

Table 6 classification table for the multinomial logistic regression

2. Binary Logistic Regression

Binary logistic regression was used to compare just two categories: the normal tissue group and the dysplastic tissue group. Following the same method as outlined previously, we arrived at table 7 for the pair wise comparisons for normal vs dysplastic tissue which again consist of the estimated coefficients (B), corresponding standard errors, *p-values* and the estimated odds ratios (Exp(B)).

Parameter Estimates

pathbinary[a]	B	Std. Error	p-value	Exp(B)
0 Intercept	-.546	.384	.155	
A2std	.532	.396	.180	1.702
A3std	.917	.358	.011	2.503
A7std	1.728	1.011	.088	5.627
A9std	1.105	.541	.041	3.021
W1std	-1.817	.741	.014	.162
W4std	2.471	1.115	.027	11.833
W5std	2.000	.819	.015	7.391
W6std	-1.862	.746	.013	.155
W8std	-.852	.543	.117	.427

a. The reference category is: 1.

Table 7 Estimated effects for the comparison of normal tissue vs dysplastic tissue

The interpretation of B and exp(B) is the same as before, except that now you only have one pair to compare. The comparison is given in table 8.

Classification

Observed	Predicted		
	0	1	Percent Correct
0	41	9	82.0%
1	10	34	77.3%
Overall Percentage	54.3%	45.7%	79.8%

Table 8 classification table for the binary logistic regression