National Physical Laboratory

NPL REPORT MS 35

WHAT AM I MEASURING? USING EXPERIMENTAL DESIGN TO UNDERSTAND AND IMPROVE MEASUREMENT PIPELINES, WITH AN EXAMPLE APPLICATION TO RADIOMICS MEASUREMENTS.

IGNACIO X. PARTARRIEU, NADIA SMITH AND PETER HARRIS

JANUARY 2022

Department for
Business, Energy
& Industrial Strategy

FUNDED BY BEIS

What am I measuring? Using experimental design to understand and improve measurement pipelines, with an example application to radiomics measurements.

Ignacio X. Partarrieu, Nadia Smith and Peter Harris
Data Science

ABSTRACT

Measurement pipelines are becoming more and more complex as technology improves. Additionally, competition between measurement providers means that technological specifications may differ significantly between instruments designed to be measuring the same underlying measurand at a similar resolution. These trends have been identified as contributing to the reproducibility challenge, as often it is difficult to establish whether measurements are different due to differences in the measurand or in the measurement pipeline. Here, we discuss a method for determining the relative sensitivity of a measurement to changes in the underlying measurand and to changes in factors defining the pipeline. We apply the method to radiomics, which is a machine learning methodology used in medical imaging to calculate a large number of image features from patient scans, in an attempt to extract information that might be of prognostic or predictive value. We show that our method can be used to give a clear prioritisation order for standardisation of those factors by quantifying their effect on the measurement. The method can also be used to understand the comparability between radiomic studies, as it makes it possible to determine whether the differences in factor levels will affect a measurand. Importantly, the method can determine the effect of the interactions of factors effecting the measurement.

An example MATLAB live script implementing a basic sensitivity analysis using the methods in this report can be made available upon request using the following contact form:

https://www.npl.co.uk/research/data-science/data-science-contact-form.

ISSN 1754-2960

https://doi.org/10.47120/npl.MS35

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

This work was funded by the UK Government's Department for Business, Energy and Industrial Strategy (BEIS) through the UK's National Measurement System programmes.

Approved on behalf of NPLML by
Louise Wright, Head of Digital Metrology

CONTENTS

GLOSSARY/ABBREVIATIONS

ANOVA:	Analysis of Variance
AU:	Arbitrary Units
Balanced design:	an experiment in which the number of data points in each combination of factor levels is equal
CAD:	Computer Aided Design
Design:	the factors and levels being investigated in an experiment
GATE:	GEANT4 Application for Emission Tomography
GLRLM:	Grey Level Run Length Matrix
Factor:	an element of a measurement pipeline being investigated which may impact our measurement result
FWHM:	Full Width Half Maximum
Level:	the ‘values’ taken by the factors, which are not necessarily numerical
Measurand:	quantity intended to be measured
PET:	Positron Emission Tomography
OSEM:	Ordered Subset Expectation Maximisation
RE:	Run Entropy

EXECUTIVE SUMMARY

1 INTRODUCTION

Measurement pipelines, i.e. the series of steps used to transform input knowledge to a measurement result (Figure 1), have become increasingly complex as knowledge advances. This increased complexity can lead to the introduction of additional uncertainties into a measurement, which need to be understood if we are to trust and use the measurement results. This requirement is especially true when factors in the measurement pipeline can be varied, as these variations may cause changes in the results that are not related to changes in the underlying quantity or property that we wish to measure.

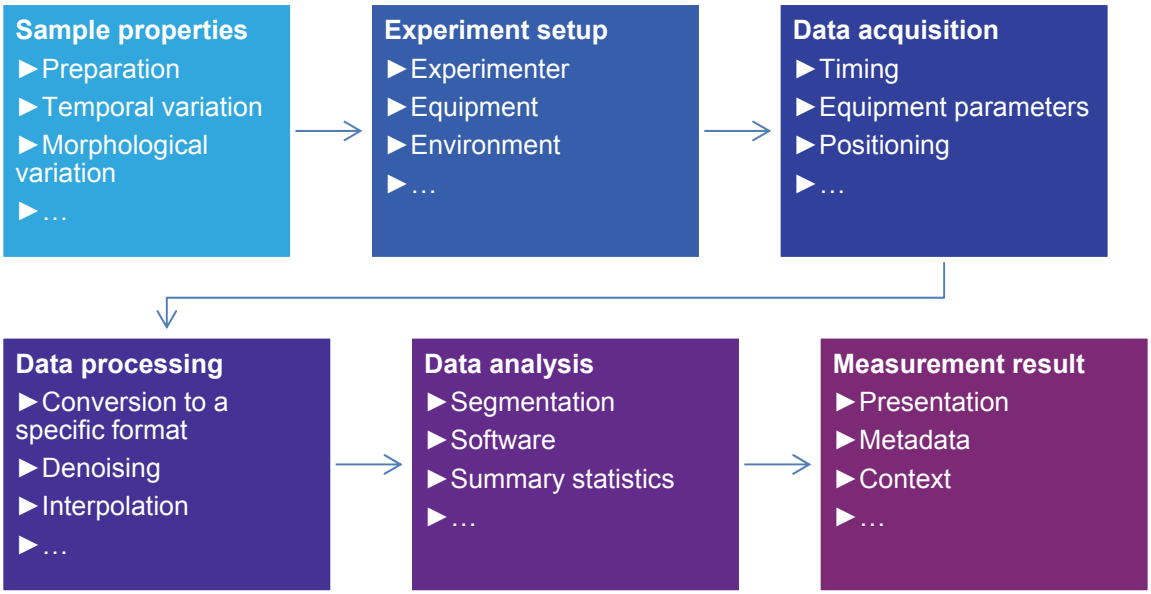

Figure 1: An example generic measurement pipeline with six major factors, each presented here with three example sources of associated uncertainty. This is not an exhaustive list, and different pipelines may vary significantly from this format. This figure is only intended to convey the general concept of a pipeline and the sources of uncertainty that can arise.

Methods of experimental design aim to establish the experimental conditions under which we can retrieve the most information from an experiment[1]. They do this by establishing efficient methodologies for exploring how a system responds to changes in its influencing factors. These methodologies are improvements upon the typical **'pick and try'** or '**varying one factor at a time'** approaches to designing experiments, and often reveal details about systems that have been missed by such experiments. Ronald Fisher[2] first devised them in the 1920s, and they have since become known as factorial experiments and are widely used in many industries.

This report focusses on the use of these methods for determining the sensitivity of radiomics measurements to the measurement pipeline versus the true physical property which they intend to measure. Radiomics measurements are a series of calculations[3] that can be performed on medical images, which have gained popularity in recent years. They attempt to extract quantitative predictive or prognostic information from an image which might not be evident to a clinician. This information is often assumed to be a measure of tumour heterogeneity. However, the design of experiments methods used here can be replicated to many different measurement pipelines to gain insights about them, and this is the main message conveyed in this report.

2 METHODS

2.1 DESIGN OF A FACTORIAL EXPERIMENT

In order to design a factorial experiment, we must first establish the factors in our pipeline that we wish to investigate. If, for example, we believe that factors A and B have some effect on the output of the measurement, we can decide on the levels that we deem most appropriate for them and undertake experiments defined by all possible combinations of those levels. For example, say A and B can each take two levels denoted, respectively, by '1' and '2' (note here that 1 and 2 are arbitrary labels and not numerical values). We then gather the data as seen in Table 1. A key point to note is that the data gathered in this manner covers all possible combinations of the factors and their levels. It is possible to still get substantial insight when this is not achievable through a fractional factorial design[4], however these will not be covered in this report. Additionally, each experimental condition defined by a particular combination of factor levels should have an equal number of data points in order to be of a balanced design, which is a requirement for a factorial experiment. For example, if an experiment with four measurement results has a single factor with two levels, each level must be associated with two measurement results, and any other distribution would be unbalanced. If using a deterministic simulation, only one data point is necessary per combination of factor levels.

Table 1: Example table demonstrating how to store data gathered from a factorial experiment. A key point to note is that all combinations of factors A and B and their levels are considered.

Factor A	Factor B	Measurement result
1	1	…
1	2	…
2	1	…
2	2	…

2.2 CALCULATING RELATIVE SENSITIVITY

Relative sensitivity refers to the sensitivity of a measurement to a factor in a given pipeline in proportion to its sensitivity to all other factors. A lot of insight can be gleaned from a system by looking at/visually assessing the values gathered as shown in Table 1, but we can often take the analysis further by calculating what is known as the sum of squares[5,6] in order to obtain such a relative sensitivity value. We will focus on what is known as Type III sum of squares here. This can be calculated for factors A and B (factor sum of squares, SSF), as well as their interactions (interaction sum of squares, SSI) and the dataset as a whole (total sum of squares, SST), as per the following equations:

$$SST = \sum_{n=1}^{N} (y_n - \bar{\bar{y}})^2 \tag{1}$$

where N is the number of data points (e.g. measurement results in Table 1), y_n is the data and $\bar{\bar{y}}$ is the grand mean, which is the mean of all the data points, regardless of factor grouping. For a factor with K levels, each with $M = N/K$ data points (i.e. having a balanced design), the factor sum of squares can be calculated as:

$$SSF = M \times \sum_{k=1}^{K} \left(\frac{1}{M} \sum_{m=1}^{M} y_{k,m} - \bar{\bar{y}} \right)^2 \quad (2)$$

where the expression in the brackets takes the mean of the data for a factor level and subtracts the grand mean from it. These values are then squared and summed, and finally weighted by the number of datapoints used in the calculation.

The sum of squares of interactions SSI can be calculated in a similar manner, where in the first instance we calculate the mean of the data for the combinations of factor levels of interest and subtract the grand mean from each mean calculated in this way. The results are squared, summed, and weighted by the number of data points. We then proceed to subtract the sum of squares that can be attributed to lower order effects of the self-same factors.

Either the factor or interaction sum of squares (denoted here as SSX, where the X can be replaced by an I or F as appropriate), taken over the total sum of squares, gives us a metric known as η^2:

$$\eta^2 = \frac{SSX}{SST} \quad (3)$$

which tells us the relative sensitivity of our measurement to changes in the pipeline factors investigated [7].

2.2.1 Guide to the practical implementation of relative sensitivity calculations

Often, the best way to calculate these sums of squares is through an analysis of variance (ANOVA), which results in a sum of squares value for each factor in the factorial experiment. A 'main effects' ANOVA can give the experimenter an idea of which factors are major contributors to variations in a measurement. If a 'full' model ANOVA is calculated, one can also obtain information about all interactions between the factors of interest as well. It should be noted that ANOVA assumes that the data is homoscedastic (it has a homogeneous variance), that samples are independent of each other and that factor effects are additive. However, when dealing with a balanced design ANOVA can be robust to violations of these assumptions, and in particular these assumptions do not necessarily affect our interpretation of η^2, as long as we remember that it represents the proportion of variance of a factor due to the selected level sampling and use caution. If complex behaviour is expected for a particular factor, it may be prudent to sample more levels. An exhaustive description of ANOVA and the calculations of η^2 values, as well as limitations both mathematical and computational are beyond the scope of this report, and we refer the interested reader to the following key texts[4,5,7,8]. It is sufficient here to note that many commonly used programming languages (python, MATLAB, Julia, etc…) have implementations of ANOVA that allow one to extract sum of squares values using an input format similar to that seen in Table 1. For larger matrices, shortcuts[9] can be manually implemented to avoid 'out of memory' errors. These can occur if many factors and interactions are being investigated, and depend on the format used to store the data (e.g. data stored as 'single' precision will take less memory than that stored as 'double' precision) and memory available to the computer (in MATLAB this can be established using the 'help memory' command). We would suggest that it can be advantageous to begin an analysis with a main effects ANOVA for scoping purposes before moving on to a full ANOVA to look into interactions.

If rigorous, one factor at a time experimentation has been carried out in the past on a measurand of interest, key improvements in the understanding of a pipeline can often be gleaned from second order interactions (interactions between pairs of factors) of which experimenters are often unaware, and so on (third order, fourth order and higher order).

A handy rule is that for a full factorial experiment the η^2 values of all factors and their interactions should sum up to unity. If this is not the case there are likely missing values or the experiment may be unbalanced.

2.3 SIMULATING POSITRON EMISSION TOMOGRAPHY DATA

For this report, we discuss the application of the methods described above to radiomics measurements in the positron emission tomography (PET) imaging pipeline[10]. PET is a sophisticated medical imaging technique, whereby certain chemical reactions can be observed in a quantitative manner. This observation is achieved by injecting a patient with a radiotracer specifically designed to label said reactions, such as those involved in glucose metabolism. The radiotracer will then accumulate at sites where these reactions happen, emitting pairs of photons as it decays that are detected for the purpose of reconstructing images[11]. This reconstruction is done using specialised algorithms which require setting parameters for the reconstruction. A non-exhaustive summary of the PET imaging pipeline is presented in Figure 2. The key steps are identified in the boxes and to the right of these are listed various factors which may affect the measurement outcome, with factors varied in our experiment highlighted in orange.

We used bespoke MATLAB software to simulate the acquisition of a phantom object where we varied the:

1. number of decay counts
2. number of iterations and
3. number of sub-iterations of an ordered-subset expectation maximisation (OSEM) algorithm
4. Gaussian filter full width half maximum (FWHM)
5. discretization of the image.

The phantom itself, described below, was also varied to have differing cone radius. The values taken by the simulations are given in Table 2. A full factorial experiment for these factors and levels resulted in 21 504 simulations to analyse, which can be pre-determined by multiplying together the numbers of levels for each factor. The advantage of simulating the data was the ease of sampling a full factorial space, however it should be noted that the simulations designed were fairly simplistic and so of limited representativeness of a true acquisition. For a more thorough examination specific to PET, more sophisticated simulations should be developed.

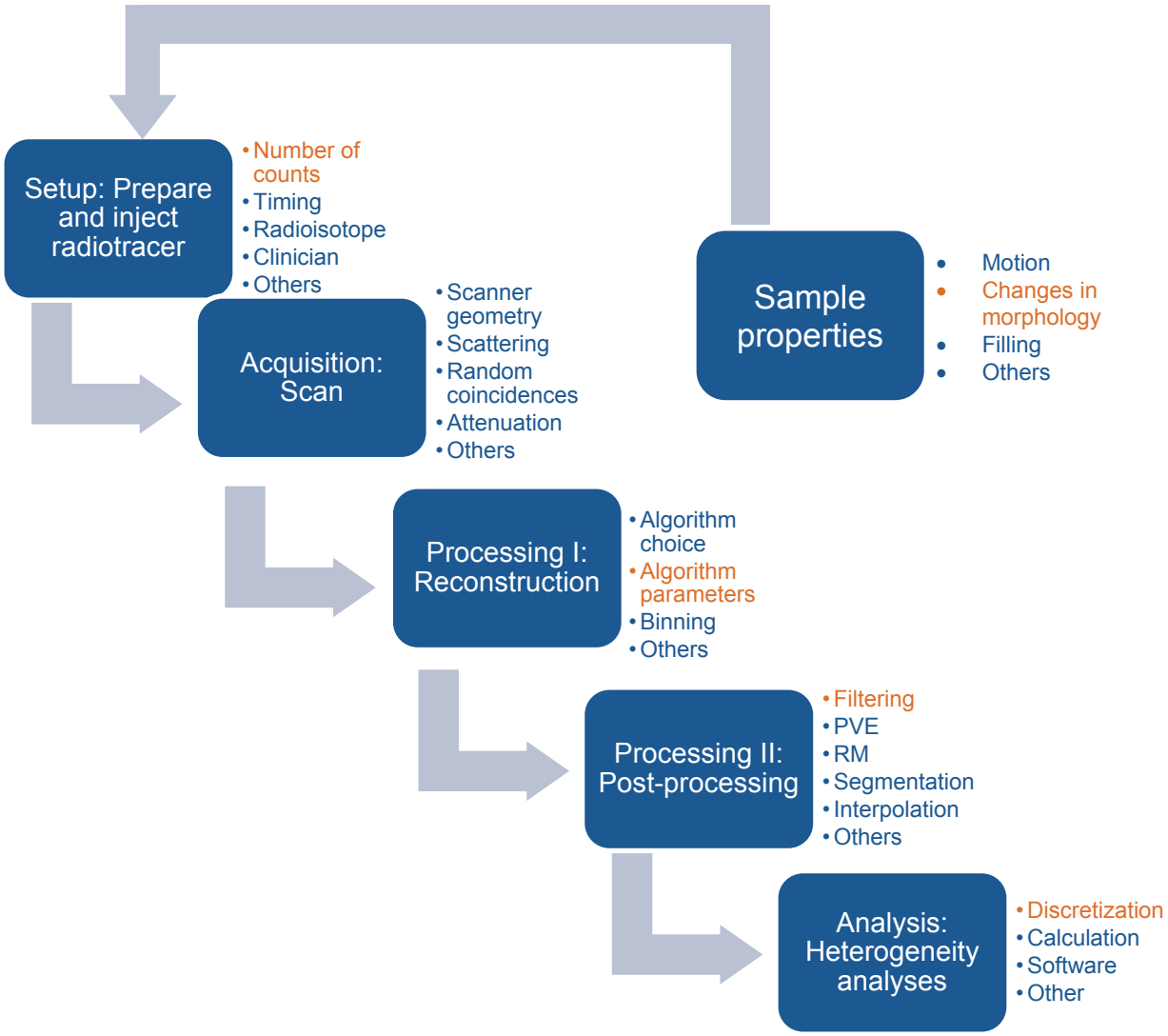

Figure 2: Example PET scanner pipeline. Selected key steps are shown in blue boxes, and associated sources of uncertainty are listed to the right of each box. The list is not exhaustive and indeed many more steps and sources of uncertainty could likely be added. The sample properties can be considered part of the pipeline, however, whereas we wish to minimise the effect on the result of other pipeline steps, we wish to maximise that of the sample properties. Text highlighted in orange represents factors varied in our experiment.

2.4 PHANTOM OBJECT

We simulated three phantom objects looking like the one seen in Figure 3 with a 100 ml volume and with varying radius for the conical intrusions. These phantoms were designed to mimic the appearance and variations of tumours measured using PET, where typically a tumour is well-perfused along the outer rim as radiotracers can accumulate there but possesses a necrotic core lacking signal (i.e. radiotracer accumulation). The conical intrusions, simulated as being filled with radiotracer, replicate this schema. The variation in the radius represent the natural heterogeneity with which tumours can occur.

Table 2: Factorial experiment design table showing factors chosen for our factorial experiment and their levels. Being balanced, each combination of factor levels in this design has an identical number of data points (n=1) associated with it.

Factor	Levels
Phantom cone radius	0.5, 1, 1.5 mm
Voxel size	2, 3, 4, 5 mm
Gaussian filter FWHM	None, 1, 2, 3, 4, 5, 6 mm
Number of iterations	4, 6, 10, 12
Number of subsets	4, 8, 16, 32
Number of counts	$10^5, 10^6, 10^7, 10^8$
Number of bins (sampling)	32, 64, 128, 256

2.5 RADIOMICS ANALYSIS

The radiomics calculations[12,13] were performed on set segmentations of the spheres, which were obtained from the binary CAD model of the phantom and known dimensions of the spheres. The radiomic metrics were calculated using the pyradiomics software package[14], and we present relative sensitivity results for all metrics for a main effects ANOVA, though we do a further full ANOVA breakdown of a specific radiomics metric to further demonstrate the advantages of the factorial experiment framework. A good radiomic metric in this case would be sensitive to changes in the cone radius, but not to variations of the other factors.

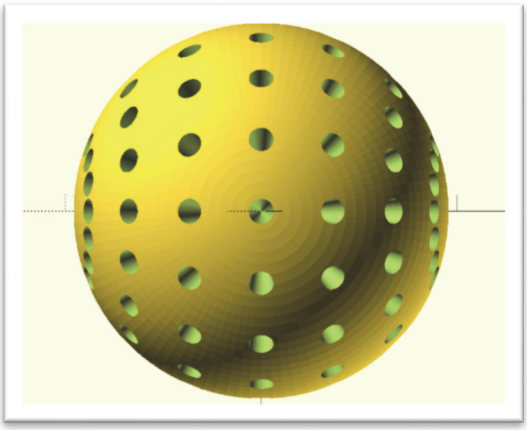

Figure 3: CAD model of a spherical phantom object with conical intrusions.

3 RESULTS

3.1 RECONSTRUCTED IMAGES

Figure 4 presents example results of two simulated reconstructions using different reconstruction parameters. The figure shows an extreme example of the variability of images, as clear differences in the structure and information available in the images may be observed.

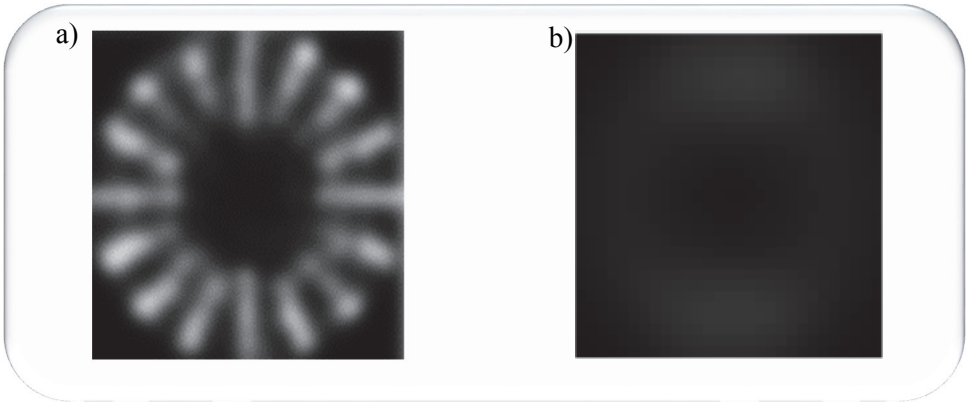

Figure 4: Examples of a) a reconstructed image with average counts and iterations, small voxels and minor filtering b) a reconstruction with low counts and iterations, small voxels and heavy filtering.

3.2 RELATIVE SENSITIVITY OF RADIOMICS METRICS TO EXPERIMENTAL FACTORS

The results of the relative sensitivity analysis are presented in Figure 5 , where the x-axis holds the various radiomics metrics calculated on the volume of interest (VOI) and the y-axis shows the relative sensitivity η^2 of said metrics to the factors examined. This plot can be used to visualise the factors that most affect the metrics calculated, which will guide us in determining which metrics are useful for the measurement of the property of interest. Specifically, as previously stated, we would want a metric to be sensitive to changes in the cone radius and not to other factors.

As such, metrics worth investigating further would be those having a large green contribution to their bar plot. Mean, maximum, total energy and grey level run length matrix run entropy (GLRLM RE) are the best performing metrics in this regard. However, all metrics are more sensitive to changes in factors and their interactions than to changes in the cone radius. Many metrics appear to be sensitive to factors such as filtering, counts and voxel size. Interestingly, metrics are not sensitive to changes in the iteration and sub-iteration parameters.

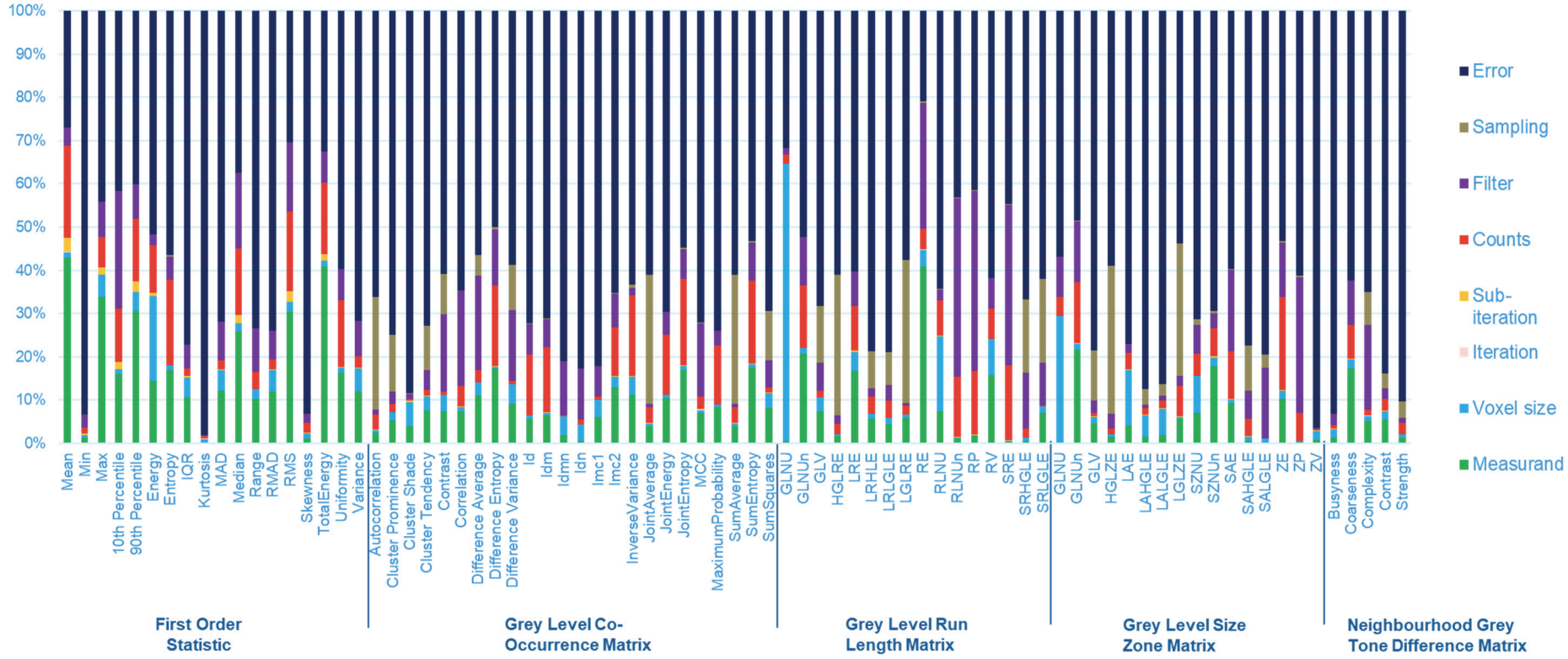

Figure 5: Relative sensitivity of radiomics metrics to changes in the factor levels. Error here denotes all interaction terms for the factors. Cone radius was the factor to which radiomic metrics should be sensitive, and so a large green bar indicates a metric which might be worth investigating further. Sensitivity to other factors can be determined from the legend.

3.3 RUN ENTROPY PARAMETER CONTRIBUTION BREAKDOWN

As GLRLM RE was one of the best performing metrics we investigated it further, breaking down contributing factors to further aid with visualisation of the analysis. The results of this analysis are the box plots seen in **Figure 6**. In these plots, we can see that the distribution of GLRLM RE values does indeed vary with respect to changes in cone radius, progressively increasing as the radius does. Similarly, it is possible to observe a large initial change with respect to the filter FWHM as simulations go from no filtering to some filtering, stabilising for FWHM values > 3 mm. The lack of sensitivity to iteration values can also be seen in these plots, with GLRLM RE values showing no visible change for the various levels of this factor. A similar breakdown could be used for other metrics if they were deemed of interest.

As GLRLM RE was particularly sensitive to filtering, and we determined that this was largely due to the change from filter to no filter, an ANOVA was run for each filter level, resulting in an associated SST for each level as well as new SSF and SSI values. The results of this analysis can be seen in Figure 7. In particular, we can see that in the case where there is no filtering, GLRLM RE is not very sensitive to changes in cone radius. However, as long as some filtering is applied the sensitivity to changes in the cone radius becomes dominant (> 60 %).

For demonstrative purposes, we also show which are the largest interaction effects for GLRLM RE once filter FWHM has been standardised. These are both related to voxel size. This is in line with expectations as GLRLM RE is sensitive to both voxel size and counts once filter FWHM has been set.

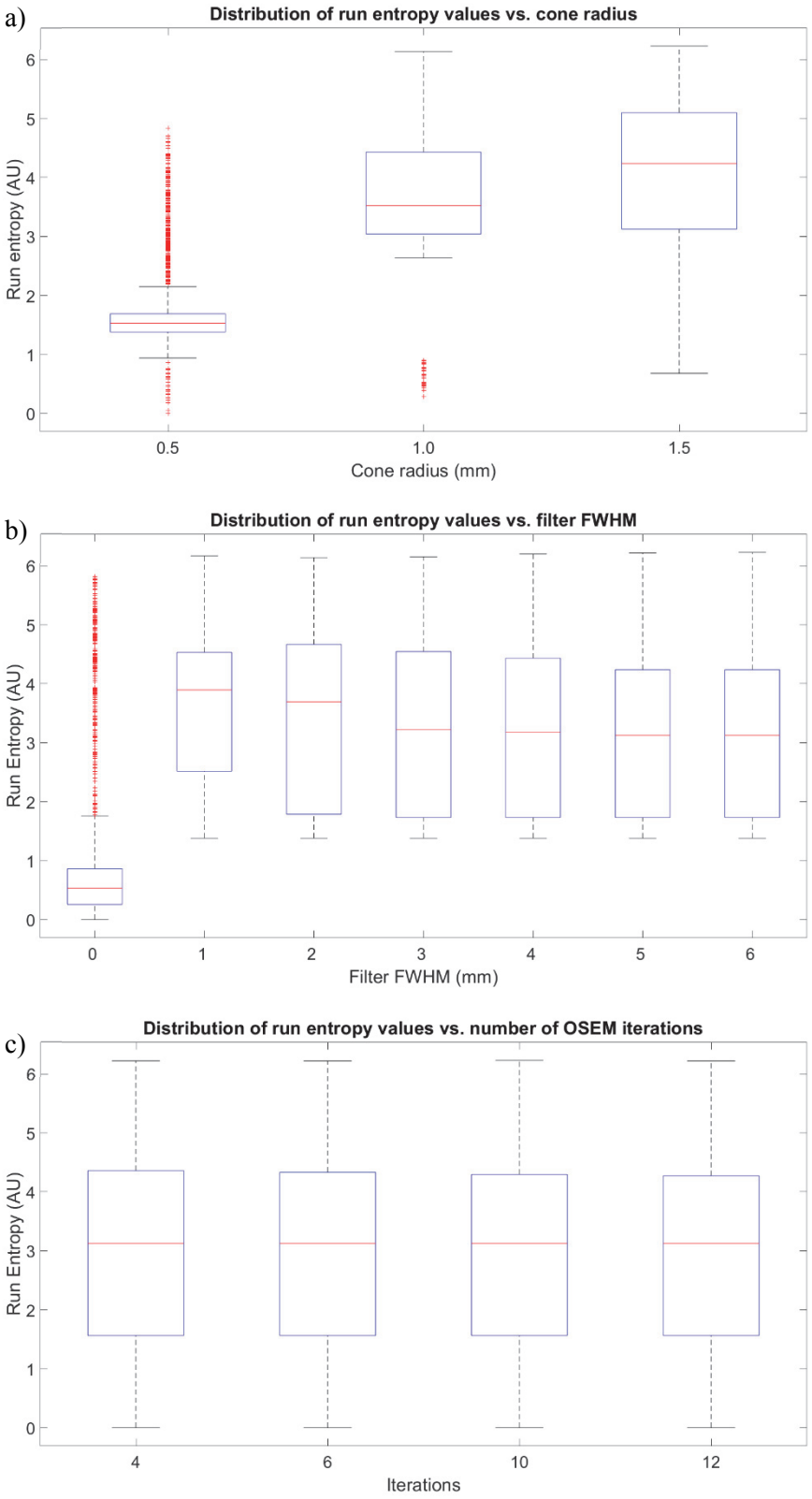

Figure 6: Box plots showing distribution of run entropy (RE) values for various factors and their levels, in order to visualise the underlying distributions which $\eta^2$is detecting. a) RE values vs. cone radius b) RE values vs. filter FWHM c) RE values vs. iterations. The red line is the median, the bottom and top of the box are the 25th and 75th percentiles respectively, the red crosses are outliers, defined as being more than 1.5 times the inter-quartile range away from the bottom or top of the box. The whiskers are the remaining values.

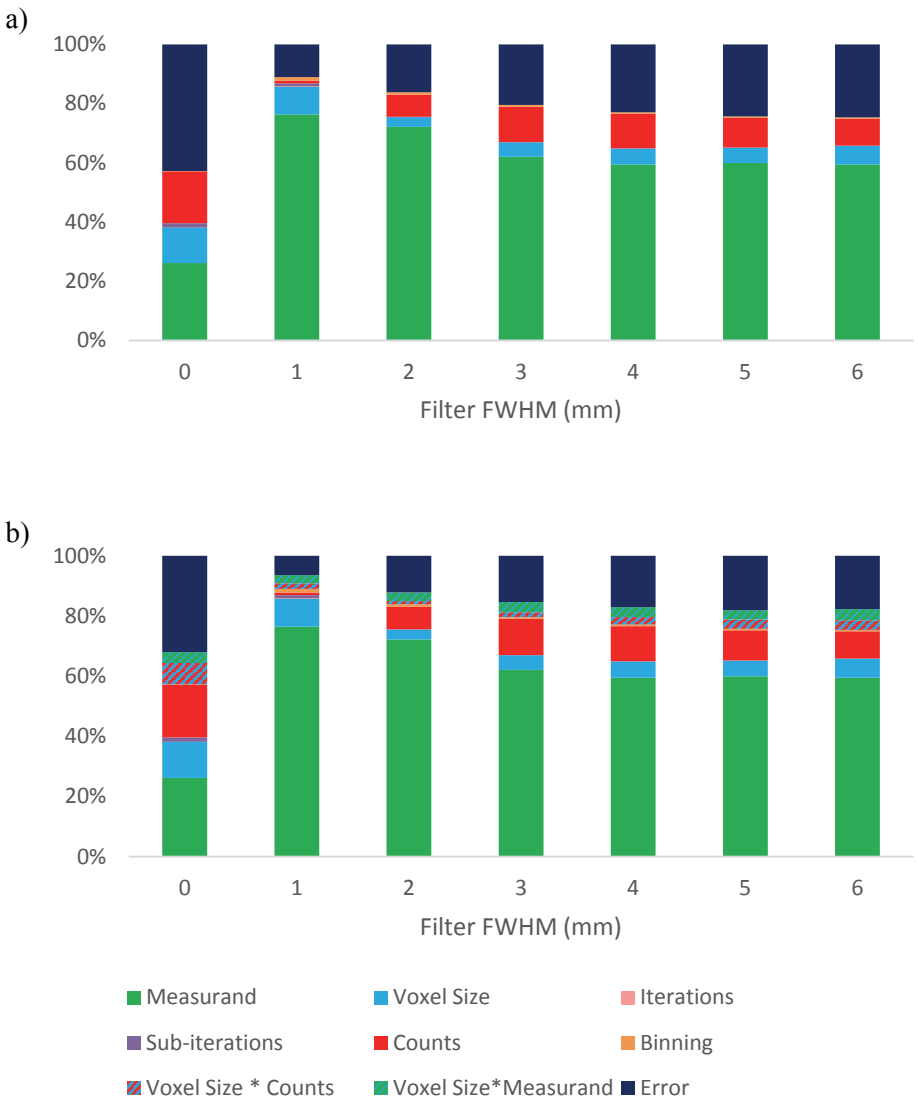

Figure 7: Relative sensitivity of GLRLM RE when filter FWHM is standardised to a) main effects of remaining factors b) main effects of remaining factors and two major interaction terms.

4 DISCUSSION

In this report we have discussed the use of factorial experiments and ANOVA to determine the relative sensitivity η^2 of measurements to factors in an imaging pipeline, with an example application to the calculation of radiomics metrics in PET. We hope to have convinced the reader of the advantages of such an approach which include but are not limited to:

- The ability to determine the relative impact of factors on measurements
    - A corollary to this being the ability to determine which factors it is most important to standardize first
- The ability to determine the measurements which are most sensitive to the factor which is of interest, and robust to those factors which are not
- The ability to detect sensitivity of measurements to interactions between factors which might not be contributing much by themselves

- The ability to establish which experimental design gives the best sensitivity to the factor of interest
- The ability of a sensitivity analysis to provide a first pass evaluation of the measurement uncertainty
- The ability to easily visualise all of the above.

This methodology can be expanded to include the concept of 'controllable' factors, which can be standardised (i.e. set to a specific level) and uncontrollable factors, which can be simulated or controlled in technical validation, but cannot be standardised. Interactions of the former are not necessarily detrimental, as once standardised their effect can be reduced or understood through appropriate level selection and analysis. However, interactions of the latter are detrimental, and if a metric assessing the measurand shows a large interaction with an uncontrollable factor it is likely unsuited to its purpose.

There are however limitations to this methodology. We have emphasized that η^2 is a measure of **relative sensitivity**. This means that the values given are a local measure of variation limited to the factor space of the experiment. So, if an additional factor was identified and investigated, we could potentially discover that it affects our interpretation of these results if said parameter has a strong effect on the measurements being made, or interacts with other factors to produce a strong effect. Two alternative measures of sensitivity, known as partial eta-squared, η_p^2 and generalized eta-squared, η_G^2, have been suggested as alternatives[12] which might be more practical for comparison between studies, however they still rely on the experimental design being similar. A rough method of understanding the reliability for η^2 values obtained for the factors investigated, if more than two levels are used for each factor, is to run the analysis multiple times leaving one of the levels out each time. We can then get an average η^2 and an associated standard deviation for the η^2 of each factor where this is done, though this assumes homoscedasticity. This can help us determine how the sensitivity might relate to the chosen factor levels.

In this report we have discussed the use of a full factorial experiment. Often, these are not possible due to cost and time constraints. Potential future work could be used to investigate sensitivity analyses for fractional factorial experiments. Additionally, the use of a fairly simplistic simulation for demonstrative purposes means that this set of simulated experiments could be assumed to have fewer errors than real-world acquisitions. It would be useful to carry out real-world acquisitions in order to understand how repeatability issues might affect the results presented here, and to use more sophisticated simulation engines such as GATE in order to simulate effects with more granularity[15,16].

Another limitation of this method is that ANOVA captures linear terms and interactions, therefore it might not accurately represent non-linear behaviours.

A point of interest that is worth noting is that the framework for variance-based sensitivity analyses has been proposed at least twice independently, resulting in differing terminologies for similar concepts in different fields[17]. First order Sobol indices S_i, defined as the variance contribution of a factor over the total variance, in particular, are in fact by definition identical to η^2, formulated as below[18,19]:

$$S_i \equiv \eta^2 \coloneqq \frac{V_{x_i}\left(E_{x\sim i}(y|x_i)\right)}{V(y)} \tag{4}$$

where $V(y)$ represents the unconditional variance of data y and $V_{x_i}\left(E_{x\sim i}(y|x_i)\right)$ represents the variance of data y for factor x and its associated levels i. A discussion of other frameworks for sensitivity analysis may be found in the NPL Report MS 2[20].

5 CONCLUSION

Factorial designs, ANOVA and sensitivity analyses are valuable methods for scientists to be aware of if they wish to determine optimal settings for their experiments, study uncertainties in pipelines and build sophisticated models which take interactions of factors into account. We have shown here some of the advantages of these methods and associated visualisation tools, using radiomics in PET as an example application. We conclude that imaging pipelines should appear as in Figure 8, with experimental design prioritised, as this will determine the amount of useful information which can be extracted from the data gathered.

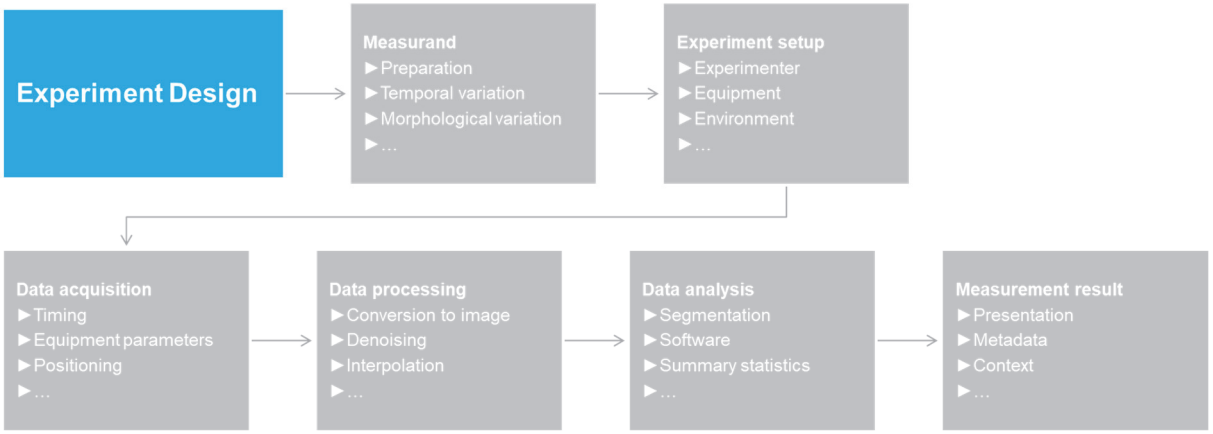

Figure 8: A corrected generic imaging pipeline, where experimental design is prioritised.

6 ACKNOWLEDGEMENTS

This work was funded by the UK Government's Department for Business, Energy and Industrial Strategy (BEIS) through the Data Science programme of the UK's National Measurement System.

I would also like to thank Daniel Deidda and Ana Denis-Bacelar for some interesting conversations around the applications in SPECT, and Louise Wright for bringing Sobol indices to my attention, as I plan to learn more about this form of sensitivity analysis. Our reviewer, Jenny Venton, also made some very helpful comments which have led to a more coherent document, and for that she has our thanks.

7 REFERENCES

1. Rojas, C. R., Welsh, J. S., Goodwin, G. C. & Feuer, A. Robust optimal experiment design for system identification. *Automatica* **43**, 993–1008 (2007).
2. Ronald A. Fischer. *Statistical Methods for Research Workers*. (Oliver and Boyd, 1925).
3. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative. *Radiology* **295**, 328–338 (2020).
4. George E. P. Box and Friends. *Improving Almost Anything: Ideas and Essays*. (Wiley, 1985).
5. Raphael Vallat. Pingouin ANOVA documentation. *Pingouin statistics* https://pingouin-stats.org/generated/pingouin.anova.html (2021).
6. Weisstein, E. W. ANOVA. https://mathworld.wolfram.com/ANOVA.html.
7. Cohen, J. Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs. *Educational and Psychological Measurement* **33**, 107–112 (1973).
8. Fisher, S. R. A. *The Design of Experiments*. (Oliver and Boyd, 1935).
9. Jeff Miller and Patricia Haden. *Statistical Analysis with The General Linear Model*.
10. Cook, G. J. R., Azad, G., Owczarczyk, K., Siddique, M. & Goh, V. Challenges and Promises of PET Radiomics. *Int J Radiat Oncol Biol Phys* **102**, 1083–1089 (2018).
11. Bailey, D. L., Townsend, D. W., Valk, P. E. & Maisey, M. N. *Positron Emission Tomography: Basic Sciences*. (Springer Science & Business Media, 2005).
12. Hatt, M., Vallieres, M., Visvikis, D. & Zwanenburg, A. IBSI: an international community radiomics standardization initiative. *Journal of Nuclear Medicine* **59**, 287–287 (2018).
13. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by non-invasive imaging using a quantitative radiomics approach. *Nat Commun* **5**, 1–9 (2014).
14. Griethuysen, J. J. M. van *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* **77**, e104–e107 (2017).
15. Jan, S. *et al.* GATE: a simulation toolkit for PET and SPECT. *Phys Med Biol* **49**, 4543–4561 (2004).
16. Sarrut, D. *et al.* A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications. *Med Phys* **41**, 064301 (2014).
17. Archer, G. E. B., Saltelli, A. & Sobol, I. M. Sensitivity measures, anova-like Techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation* **58**, 99–120 (1997).
18. Saltelli, A. *et al.* Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software* **114**, 29–39 (2019).
19. Karl Pearson. *On the General Theory of Skew Correlation and Non-linear Regression*. (1905).
20. Esward, Trevor, Matthews, C., Wright, L. & Yang, X.-S. *Sensitivity analysis, optimisation, and sampling methods applied to continuous models*. 44 (2010).