The Use of Web Techniques for Revising RCT PILs

The Use of Text Analytics & Crowdsourcing to Acquire, Measure and Analyse Public Reviews on Patient Information Leaflets for Randomized Clinical Trials.

MSc. Santos Sanchez F.

J.C.Wyatt@soton.ac.uk

Prof Wyatt J.

J.C.Wyatt@soton.ac.uk

Prof Tiropanis T.

t.tiropanis@southampton.ac.uk



## Summary

Clinical trials have become a corner stone (Lovato, Hill, Hertert, Hunninghake, & Probstfield, 1997) for identifying high-quality interventions in the health-care systems of developed countries. They enable researchers to compare the effects of new drugs and treatments against those that are currently employed, to improve the health-care of the general population by developing new guidelines and practices (NHS, 2017). On the other hand, their very nature implies a risk for the patients who choose to participate, of either receiving a sub-optimal treatment or suffering previously undiscovered side-effects (Moore & Savage, 2002). Thus, it is of great importance to ensure the patient is aware of the risks and to enforce ethical practice during recruitment (MRC, 2016).

Therefore, the development of Patient Information Leaflets (PILs) which are able to inform patients about essential trial features is one of the core tasks for any UK clinical trial. The current process for clinical research is based on the NHS proportionate approach to consent (HRA, 2017), which enables most PILs for Randomised Controlled Trials (RCTs) to be designed by filling on template forms provided by the HRA and be reviewed by an Ethics Panel as part of the research submission. Nonetheless, even if this information is recognized as an essential part of any RCT by the HRA (NHS, 2017), several independent studies in the last decade have consistently found most PILs have serious issues on informing patients despite fulfilling the legal requirements and following NHS recommended guidelines and templates (Reinert, et al., 2014) (Gillies, Huang, Skea, Brehaut, & Cotton, 2014) (Poplas-Susíc, Klemenc-Ketis, Kersnik, & others, 2014) (Knapp, Raynor, Silcock, & Parkinson, 2011) (Nicholls, Hankins, Hooley, & Smith, 2009). Several different approaches have been sought to address these issues from employing quantitative content analysis of the PILs’ text to engaging with Patient and Public Involvement (PPI) groups. However, these topics have remained a research priority as evidenced by The Health Research Board Trials Methodology and Networks (TMRN) work with the James Lind Alliance and the TrialForge to setting priorities for trial recruitment research (Healy, et al., 2018). Specifically, identifying which information should be communicated to patients, assessing the effect of PPI collaboration on recruitment rates and finding the best methods to deliver information are among the top five questions identified by this JLA priority-setting panel (Healy, et al., 2018).

Therefore, it is our intention to build a tool that supports the quantitative assessment and comparison of several techniques effects on PIL readability and ease of understanding of essential trial features. Our work explores the use of textual characteristics of PILs for clinical trials, using thematic analysis on PIL comments from a PPI group, employing quantitative readability metrics to identify sentences that are too hard to understand by general audiences, the use of quantitative procedures to assess the readability of the document and the health literacy skill of the readers, employing Amazon crowdsourcing to rewrite sentences that were deemed too hard to understand, the use of a Web platform to collect, link and present the data generated during the revision process by members of the public.

## Preface

### Contributions & Acknowledgements

I dedicate this work to my father who taught me to think for myself and question the world, and whose unconditional support made possible the realization of this dream. I also wish to thank my mother and siblings whom I have dearly miss every day of this five years, and my supervisors who have been with me when the though parts have come and I had lost my motivation. In addition, I wish to recognize the support of the Mexican government and the University of Southampton for providing the necessary economic funds for this research.

### Scope

This thesis has been organized in three general parts describing the research process to consolidate the different approaches into creating PILs that are easier to understand by public audiences by employing Web techniques. In the first part of the thesis we focus our research into assessing the essential characteristics of the PILs’ texts, determining their emotive composition and the feasibility of employing diverse techniques to assess their contents. In the second part of the thesis we explore the themes and composition of public comments given to PILs with low readability and poor recruitment rates. In the final part of the thesis we evaluate the feasibility of both employing a Web platform to collect, associate, analyse and present public feedback on PILs and the use of crowdsourcing to revise PILs sentences that are deemed too hard to understand.

### Publication

PhD Symposium Paper: A Web Platform for Public Involvement Reviewing Patient Information Leaflets for Randomized Clinical Trials in the UK.

PhD Symposium Paper: Analysis of Public Comments on Patient Information Leaflets as a Measure of Quality Perception and Patient Understanding.

Unpublished Full Paper (16 pages): Analysing Public Feedback on Patient Information Leaflets: Reviewer Perception of Quality and Objective Understanding of the Information.

## Table of Contents

[Summary 2](#_Toc21714916)

[Preface 3](#_Toc21714917)

[Contributions & Acknowledgements 3](#_Toc21714918)

[Scope 3](#_Toc21714919)

[Publication 3](#_Toc21714920)

[Table of Contents 4](#_Toc21714921)

[Table of Figures 6](#_Toc21714922)

[Chapter 1: Patient Information Leaflets for Randomized Clinical Trials 7](#_Toc21714923)

[The importance of Patient Information PILs 8](#_Toc21714924)

[Procedure for Designing PILs for RCTs in the UK 9](#_Toc21714925)

[PIL Quality Issues 9](#_Toc21714926)

[Chapter 2: Methods for Increasing the Readability and Understandability of Textual Information 11](#_Toc21714927)

[NHS Proportionate Approach to the Process of Seeking Consent 11](#_Toc21714928)

[HRA PIL Guidance & the Use of Templates 11](#_Toc21714929)

[Employing Patient & Public Involvement Groups 11](#_Toc21714930)

[Quantitative Analysis of Text 11](#_Toc21714931)

[Readability Indexes 11](#_Toc21714932)

[Web Text Analysis 11](#_Toc21714933)

[Analysis of Sentiment and Emotion in Text 12](#_Toc21714934)

[Maximization of Information Delivery a Psychological approach to Leaflet Design 13](#_Toc21714935)

[Clustering techniques for identifying core characteristics 17](#_Toc21714936)

[Chapter 3: Analysis of the characteristics of PILs text 18](#_Toc21714937)

[Research Questions 18](#_Toc21714938)

[Aims 18](#_Toc21714939)

[Justification 18](#_Toc21714940)

[Background & Rationale 19](#_Toc21714941)

[Methods 19](#_Toc21714942)

[Sample & Recruitment 19](#_Toc21714943)

[Limitations 20](#_Toc21714944)

[Procedure 20](#_Toc21714945)

[Measures & Analysis 21](#_Toc21714946)

[Results 22](#_Toc21714947)

[Discussion 27](#_Toc21714948)

[Strengths & Weaknesses 27](#_Toc21714949)

[Chapter 4: Analysing Public Comments on PILs with Low Readability and Poor Recruitment Rates. 28](#_Toc21714950)

[Summary 28](#_Toc21714951)

[Research Questions 28](#_Toc21714952)

[Aims 28](#_Toc21714953)

[Background & Rationale 28](#_Toc21714954)

[Methods 29](#_Toc21714955)

[Sample 30](#_Toc21714956)

[Recruitment 30](#_Toc21714957)

[Procedure 30](#_Toc21714958)

[Analysis & Metrics 32](#_Toc21714959)

[Results 35](#_Toc21714960)

[Sample detail 35](#_Toc21714961)

[Comment analysis: How many reviewers are needed to obtain a good coverage of comments? 36](#_Toc21714962)

[Comment analysis: Classification and coding 40](#_Toc21714963)

[Association between the percentage of correct answers, PIL quality grade and the comments provided by the reviewers. 53](#_Toc21714964)

[Demographic factors associated with the PIL quality grades 55](#_Toc21714965)

[Demographic factors associated with the reviewers’ understanding of trial features 59](#_Toc21714966)

[Validation analysis: Lexicon polarity vs manual annotation 61](#_Toc21714967)

[Conclusions 63](#_Toc21714968)

[Chapter 5: Assessing the feasibility of using a Web Platform and Amazon Crowdsourcing to Revise the Text of RCT PILs. 64](#_Toc21714969)

[Research Questions 64](#_Toc21714970)

[Aims 65](#_Toc21714971)

[Background & Rationale 65](#_Toc21714972)

[Methods 66](#_Toc21714973)

[Employing a Web platform to collect public feedback on PIL text. 66](#_Toc21714974)

[Using the Cloze procedure to identify jargon terms and structural problems in the PIL sentences. 66](#_Toc21714975)

[Identifying the required health literacy skill needed to revise PIL sentences 66](#_Toc21714976)

[Assessing the learning and fatigue effects when revising PILs sentences 66](#_Toc21714977)

[Assessing the viability of the proposed revisions under the PIL authors perspectives 66](#_Toc21714978)

[Chapter 6: Research Schedule 67](#_Toc21714979)

[Bibliography 67](#_Toc21714980)

## Table of Figures

[Figure 1 Plutchik’s Wheel of Emotion (Plutchik, 1984) 13](#_Toc21715119)

[Figure 2 The components of Memory (Jansson-Boyd, 2010) 14](#_Toc21715120)

[Figure 3 The elements of learning (Jansson-Boyd, 2010) 16](#_Toc21715121)

[Figure 4 Average number of words in news articles and PILs 22](#_Toc21715122)

[Figure 5 Average percentage of lexicon words identified as emotional 23](#_Toc21715123)

[Figure 6 Comparison of PILs with news articles from 3 sources 24](#_Toc21715124)

[Figure 7 Comparison of PILs recruiting well (at least 80% of planned sample) and badly 24](#_Toc21715125)

[Figure 8 Percentage of words related to each emotion category in the news articles & PILs 25](#_Toc21715126)

[Figure 9 Percentage of words related to Positive and Negative sentiments in the control documents & PILs 26](#_Toc21715127)

[Figure 10 Percentage of words related to each emotion category in love letters, hate mail, suicide notes & PILs (MSc research) 26](#_Toc21715128)

[Figure 11 Comparison of the number of general and specific comments given to each PIL 36](#_Toc21715129)

[Figure 12 Contribution of specific comments per reviewer 37](#_Toc21715130)

[Figure 13 Distribution of specific comments taking the full sample of reviewers 37](file:///C:\Users\Fernando\Dropbox\PHD%20THESIS\FoM%20Reports\3rdViva\Thesisv0.9.docx#_Toc21715131)

[Figure 14 Distribution of comments when excluding reviewers who did not gave any specific comment 38](#_Toc21715132)

[Figure 15 Association between PIL quality grade and the percentage of correct EQIP answers 54](#_Toc21715133)

[Figure 16 Association between the number of comments on specific sections and the percentage of correct EQIP answers obtained by the reviewer 55](#_Toc21715134)

[Figure 17 Ordinal demographic factors correlated with the PIL quality grade 56](#_Toc21715135)

[Figure 18 [Boxplot] Quality grades grouped by interest in participating in an RCT 57](#_Toc21715136)

[Figure 19 [Boxplot] Quality grades grouped by gender 57](#_Toc21715137)

[Figure 20 [Boxplot] Quality grades grouped by origin 58](#_Toc21715138)

[Figure 21 Ordinal demographic variables associated with the percentage of correct EQIP answers 59](#_Toc21715139)

[Figure 22 [Boxplot] Percentage of correct EQIP answers grouped by gender 60](#_Toc21715140)

[Figure 23 [Boxplot] Percentage of correct EQIP answers grouped by origin 60](#_Toc21715141)

[Figure 24 [Boxplot] Percentage of correct EQIP answers grouped by interest in participating in an RCT 60](#_Toc21715142)

# Patient Information Leaflets for Randomized Clinical Trials

The development of PILs to inform patients about essential trial features is one of the core tasks for any clinical trial run in the UK. This information is commonly presented as patient information PILs, sheets, online documents or videos that complement or enhance the explanations given by the trial recruiters. Currently, these documents are regulated by following the pre-set formats and guidelines on the best practice for medical research set by the Health Research Authority (HRA). Under these guidelines the PILs must have an impact on the participants’ decision if they are to accomplish their primary goal:

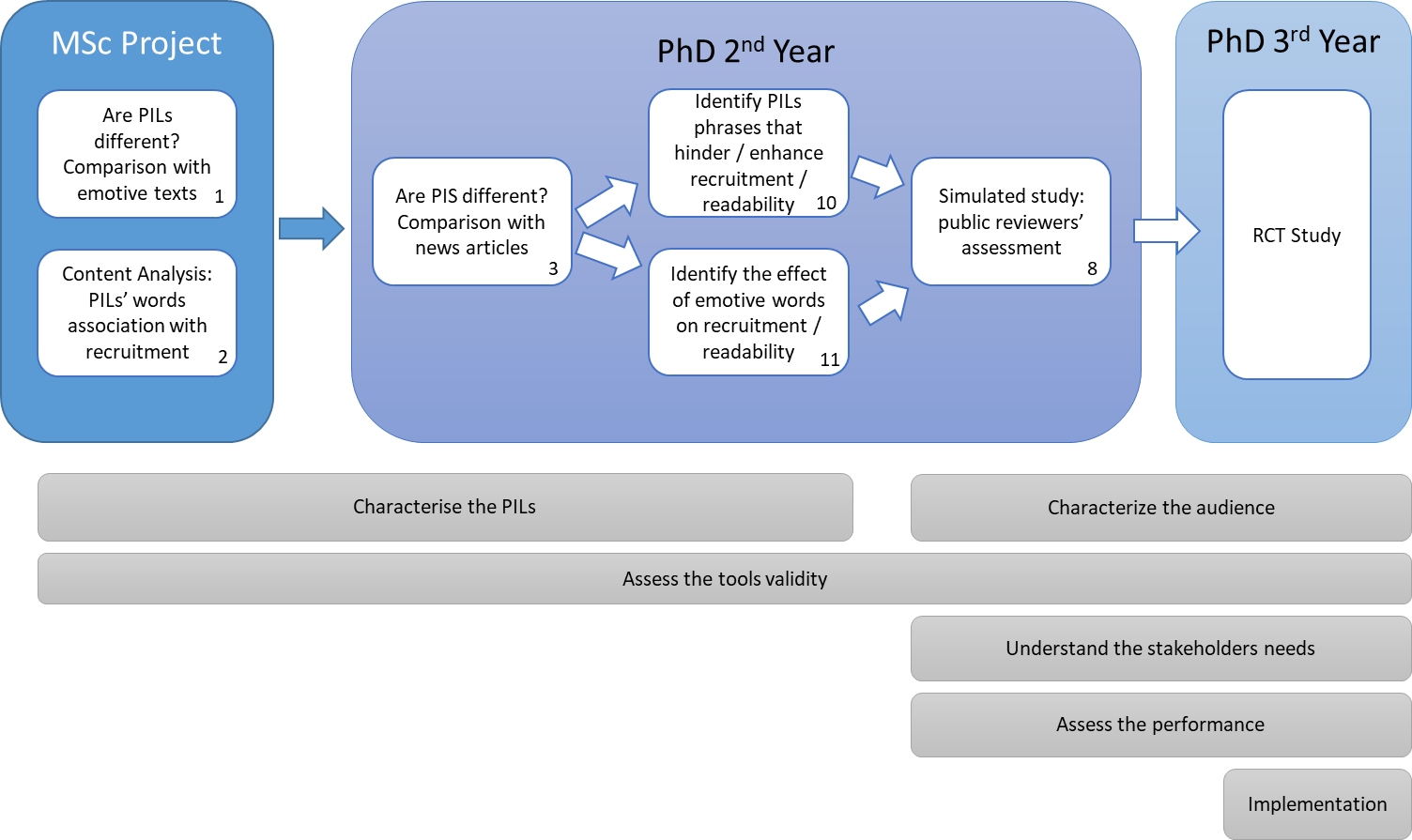
“The Participant Information Sheet should support the consent process by helping to ensure that all those who are invited to take part in a research study have been adequately informed” and “should enable potential participants to make an appropriate decision that is right for them” - (MRC, 2016) (MHRA, 2016).

Despite official recognition of the importance of these documents (NHS, 2017), several concerns have been risen about their quality in the last decade. The lack of a rigorous method for assessing the quality of written patient information, materials that are difficult to read (Moult, Franck, & Brady, 2004), inaccurate content (Moult, Franck, & Brady, 2004) (Nicholls, Hankins, Hooley, & Smith, 2009) (Escudero-Carretero, et al., 2013) and insufficient quality on most evaluated categories (e.g. text length, legibility, layout, visual structure) except ethical and legal requirements (Reinert, et al., 2014) are high priority research topics of the BRM-TMRN 2016 study (Healy, et al., 2018).

On the other hand, the fast retrieval, processing and analysis of massive amounts of text have become core activities in our current Web model. These tasks are commonly called as text (data) mining and the set of techniques employed to model and structure the information is referred as text analytics (Association, 2007), (Grimes, 2007). Furthermore, the inherent challenges of working with unstructured data formats have been recognized since the late 50’s:

"...utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the ’action points’ in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points." - H.P. Luhn, October 1958 IBM Journal article

This has created many techniques in areas like information retrieval, named entity recognition, disambiguation, co-reference, relationship and content analysis that could reveal valuable insights when applied to the PILs. In this project, we explore the effects of adding an information retrieval system (for previous PPI comments) and content analysis reports to the feedback normally given by public reviewers when reviewing PILs for low risk trials. The insights that cluster analysis can provide about the inherent relationships present in the documents, employing readability metrics to objectively quantify the difficulty of the documents and using sentiment analysis to detect the opinions and perceptions of the reviewers could greatly enhance the feedback given to a PI designing a new PIL.



## The importance of Patient Information PILs

I have previously mentioned that providing the patient with information to make an informed decision is a fundamental part of trials in the UK (NHS, 2017). This information generally includes Patient Information PILs, sheets and documents, which “should support the consent process by helping to ensure that all those who are invited to take part in a research study have been adequately informed” and “should enable potential participants to make an appropriate decision that is right for them” (NHS, 2017), MRC. These elements can form a baseline for the different interventions that should be associated with recruitment to RCTs. However, “Despite the recent focus on improving the quality of patient information, there is no rigorous method of assessing quality of written patient information” (Moult, Franck, & Brady, 2004). As has been previously commented, the HRA guidelines encourage the researchers to employ heavily standardized forms and formats with only general advice given in how to describe the RCT consist of considering the “intended audience”, employing “clear language” and to involve potential patients in the drafting of the PIL (MHRA, 2016). This has created a widespread view in the clinical community that the PILs must employ “everyday language” and explain complex words and clinical jargon but employ a “respectful tone” (Charvet-Berard, Chopard, & Perneger, 2008). The following sections explore the general literature on the aspects approached by our research proposal.

## Procedure for Designing PILs for RCTs in the UK

To be added

## PIL Quality Issues

Most of the recent research on PILs have focused on determining their quality or developing an objective method of measuring their quality, in response to Moult (Moult, Franck, & Brady, 2004). These studies have commonly found that the quality of the PILs is not optimal, often requiring a higher reading age than recommended and containing inaccuracies (Moult, Franck, & Brady, 2004) (Nicholls, Hankins, Hooley, & Smith, 2009) (Escudero-Carretero, et al., 2013). It is also a common perception in different research stakeholders (recruiters, nurses, doctors, researchers and ethic committee members) that PILs have no actual influence on the patient decision to participate and are in most cases not read or remembered (Poplas-Susíc, Klemenc-Ketis, Kersnik, & others, 2014). This brings into question if the PILs are fulfilling their role of supporting the patient decision-making process, as detailed by UK clinical regulations (NHS, 2017). This section explores some of the most commonly employed methods to assess PIL quality.

The most common assessment criteria to evaluate the quality of PILs are readability metrics, which are employed by virtually all the studies in the area in one form or another (Reid, et al., 1995) (Knapp, Raynor, Silcock, & Parkinson, 2011) (Escudero-Carretero, et al., 2013) (Gillies, Huang, Skea, Brehaut, & Cotton, 2014) (Reinert, et al., 2014). The particular metrics selected by each study vary from simple measurement of length (in either words or pages) or font size (Knapp, Raynor, Silcock, & Parkinson, 2011) to the employment of specialized formulas and instruments like the Flesch-Kincaid (Gillies, Huang, Skea, Brehaut, & Cotton, 2014) or Flesch-Formel (Reinert, et al., 2014) coefficients and the SMOG/INFLESZ scores (Escudero-Carretero, et al., 2013). In addition, Knapp (Knapp, Raynor, Silcock, & Parkinson, 2011) carried out qualitative work to measure reading times, interest in the topics, and comprehension of the topics. The readability results of these studies were similar in all cases, concluding that the PILs required higher reading skills than those recommended by the guidelines (Nicholls, Hankins, Hooley, & Smith, 2009) (Gillies, Huang, Skea, Brehaut, & Cotton, 2014) (Reinert, et al., 2014). Reinert’s study on neuro-oncology phase III trial PILs (Reinert, et al., 2014) determined that five of the nine PILs analysed required graduate levels to be read and understood.

Other characteristics employed by Reinert to determine the quality of the PILs were the page layout, and evaluations of the ethical and legal requirements, and scientific and social evidence (Reinert, et al., 2014). For the evaluation of the layout, four aspects were considered: the use of subheadings, correspondence between the heading topics and subheadings, the inclusion of a study process flow-chart and the quality of tables and illustrations. According to Reinert, evaluation of the ethical and legal requirements was done by employing a checklist for informed consent created by Harnischmacher. A questionnaire was created to assess the social evidence (PIL provides answers to patients’ frequently asked questions) based on selected items on the Patients’ Frequently Asked Questions, while the assessment of scientific evidence was done in accordance to the DISCERN criteria (Reinert, et al., 2014). Finally, Gillies’ study employed qualitative analysis to assess the degree of support that the PILs provide to the patients decision-making process (Gillies, Huang, Skea, Brehaut, & Cotton, 2014)**.**

The results provided by these studies were uniform across all authors. The patient information PILs, sheets and documents were suffering from severe deficiencies in their quality, which could affect their role in supporting patients to make a decision. Nicholls’ survey on PILs for skin cancer found that all but one PIL required education above primary level. A qualitative study on drug PILs (Poplas-Susíc, Klemenc-Ketis, Kersnik, & others, 2014) determined that the patients do not read the full PILs and consider the language too scientific.

An RCT to evaluate the use of user testing in the design of a PIL (Knapp, Raynor, Silcock, & Parkinson, 2011), found that current patient information sheets are not fit for purpose and may not have enabled valid consent by evaluating the ability of the readers to find and understand facts. Knapp also found that employing user testing could dramatically improve the quality of the PIL: “66% who read the revised PIL showed understanding of all aspects, compared to 15% of those who read the original” (Knapp, Raynor, Silcock, & Parkinson, 2011).

Reinert’s results show that “All patient informed consent documents were of insufficient quality in all categories except that ethical and legal requirements were fulfilled” (Reinert, et al., 2014), and hypothesises that there may exist a conflict between the need to inform about technical details, employ basic language and the legal requirements when designing a PIL. These observations are supported by Gillies study that found the PILs provided for trials on UK Clinical Trial Unit websites did not support good quality decision-making (Gillies, Huang, Skea, Brehaut, & Cotton, 2014).

# Methods for Increasing the Readability and Understandability of Textual Information

## NHS Proportionate Approach to the Process of Seeking Consent

To be Added

### HRA PIL Guidance & the Use of Templates

To be Added

### Employing Patient & Public Involvement Groups

To be Added

## Quantitative Analysis of Text

### Readability Indexes

### Web Text Analysis

The Web has become one of the most powerful tools invented by man mainly because the development of methods to analyse huge amounts of data and find the most relevant results to a query. To do this a new area has been created: Web Analytics “*the measurement, collection, analysis and reporting of web data for purposes of understanding and optimizing web usage*”(Web Analytics Association, 2007). This project focuses on a subgroup of methods, metrics and analysis techniques called content analysis. Content analysis has been developed to quantitatively assess the content of text resources by systematically evaluating texts to deliver replicable and valid inferences (Duriau, Reger, & Pfarrer, 2007).

Clustering and sentiment analysis methodologies appear appropriate to evaluate the PILs. Clustering is the idea of grouping together individuals, objects or elements that are more similar to themselves than they are to the elements outside the set (Clements, 1954). Sentiment analysis corresponds to the quantification of emotion related words present in a text, this analysis is commonly divided into two areas: the analysis of emotional states (e.g. anger, joy, sadness) and the study of sentiment polarity (positive, negative and neutral) (Qu, Shanahan, & Wiebe, 2004). Both of these methodologies have been extensively employed in the analysis of Web text, as we will explore in the following sections.

### Analysis of Sentiment and Emotion in Text

Sentiment analysis can be defined as the process of categorizing the polarity of a text (Qu, Shanahan, & Wiebe, 2004) while polarity would refer to the attitude of the text in the positive, negative and sometimes the neutral scale (Stone, Dunphy, Smith, & Ogilvie, 1968). This process systematically identifies, extracts, quantifies and studies affective states and subjective information present in a text (Volcani & Fogel, 2006) and can be automated to evaluate attitude as:

* Judgements/evaluations: Assess the overall perception of a product/topic in an audience
* Affective states: Identify the emotional state of the author
* Intended emotional communication: Evaluates the intended emotional effect of the document

In this research, Sentiment Analysis is employed to evaluate the polarity of PILs on the negative-positive scale and the presence of emotional words in the documents. The characterization of documents based in their emotional content would help during the analysis process detailed in the following chapters to evaluate the effect the PILs have in the patients’ decision.

To do this, the NRC Emotion Lexicon version 0.92 (Mohammad & Turney, NRC emotion lexicon, 2013) that contains the emotional relationships for 14,245 words has been employed. An emotion lexicon is a list of words and their relationships with emotion categories, these relationships in the NRC Lexicon correspond to the central emotion categories on the Plutchik’s Wheel of Emotion (Plutchik, 1984) and to the positive-negative sentiment categories.

Plutchnik’s model of emotion is based in the psycho-physiological models created by Darwin (Plutchik, 1984). Darwin’s model (Darwin, The expression of the emotions in man and animals, New York: D, 1872) assumes that the evolution process also affects the “mind” by drawing similarities between the expression of “basic emotions” between animals and humans. In accordance with Darwin’s theory, basic emotions increase the individuals’ chance of survival by providing appropriate and fast reactions during emergency events (Darwin, The expression of the emotions in man and animals, 1998). These basic emotions would therefore be a basic component in our interpersonal interactions as they could signal our imminent actions or intentions to others (Darwin, The expression of the emotions in man and animals, 1998).



Figure Plutchik’s Wheel of Emotion (Plutchik, 1984)

Based on Darwin’s model, Plutchik recognized eight primary emotions: Anger, Anticipation, Joy, Trust, Fear, Surprise, Sadness and Disgust. These emotions could be expressed at three intensity levels, with lower levels of intensity increasing the difficulty to differentiate between them. In addition, emotions nearing each other would be more similar, while emotions in opposite positions become polar opposites. Emotions that are polar opposites induce opposite effects in the individual, e.g. fear would induce an individual into fleeing but anger would make it attack.

## Maximization of Information Delivery a Psychological approach to Leaflet Design

Classical psychology studies relate the processes of learning, attention and recall (memory) to emotional stimulus (Darwin, The expression of the emotions in man and animals, New York: D, 1872), (Darwin, The expression of the emotions in man and animals, 1998), (Ekman P. , 1973), (Ekman & Davidson, 1994), (Kreutzer & Charlesworth, 1973), (Averill, 1980), (Blaney, 1986). In addition, several applied studies in consumer psychology carried on the last few decades (Adams, 1916), (Eighmey & Sar, 2007), (Escudero-Carretero, et al., 2013), (Eysenck, 2012) (Watson), (Meyer) appear to confirm these theories. While recently there has been an increase in efforts to objectively analyse the quality of the PILs, only rarely is the “*tone of voic*e” taken into account. Even the few studies, which explore the effects of different tones of voice (Escudero-Carretero, et al., 2013), (Richard, et al., 1999) on PILs, have not carried out a quantitative evaluation. The outcome of these studies shows that different tones may affect the attention and perception of the patients (Richard, et al., 1999); however, there has not been a systematic effort to assess the magnitude of these effects nor its implications for PILs in RCTs. Furthermore, certain validated scales like the Expanded EQIP scale (Charvet-Berard, Chopard, & Perneger, 2008), (Moult, Franck, & Brady, 2004), (Nicholls, Hankins, Hooley, & Smith, 2009) employed to measure the quality of drug PILs (Nicholls, Hankins, Hooley, & Smith, 2009), consider that deviations from a “*respectful tone*” weaken the PIL quality.

Consumer psychology relates to the understanding of consumer behaviour and while it is normally related to marketing techniques, focuses more on understanding the psychological aspects of behaviour than on selling products (Jansson-Boyd, 2010). This section explores the aspects most relevant to designing effective information PILs: memory, attention and learning. These processes are classically associated with emotions in the psychological literature. In addition, an exploration the practical approaches to identify associations commonly made in the marketing of new products was deemed necessary, while relating them to the classical psychological definitions for their process. From a psychological perspective based on consumer behaviour the first process, memory, can be summarized in the following model (Jansson-Boyd, 2010):

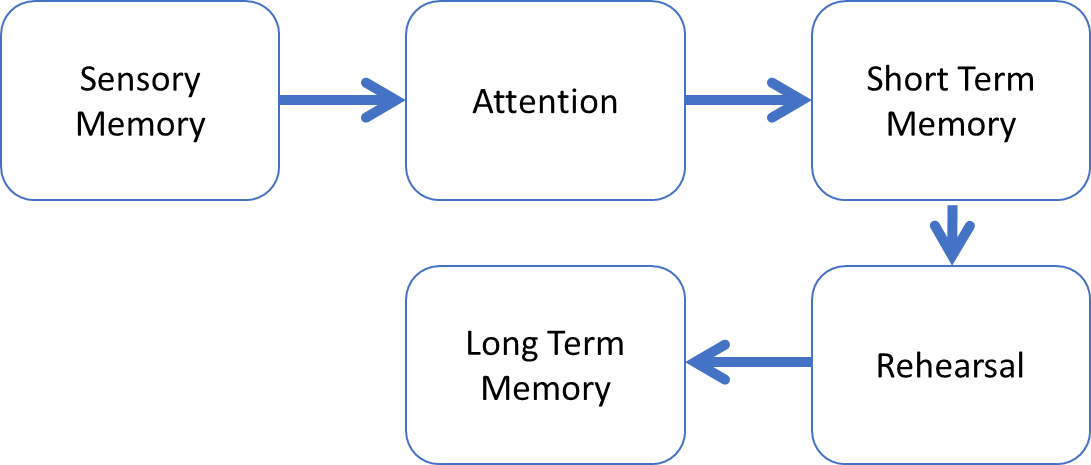


Figure The components of Memory (Jansson-Boyd, 2010)

To summarise, we constantly receive sensory information but only consciously process the subset of information that gains our attention. This information is stored in our short-term memory that is severely limited, once there it requires rehearsal to transfer the information into the long-term memory. This is particularly relevant when taking into account that emotion affects attention (Ekman P. , 1973), specifically that people in a certain mood or emotional state attend better to stimuli congruent with their emotional state (Bower, How might emotions affect learning, 1992), (Bower & Morrow, Mental models in narrative comprehension, 1990), (Ekman & Davidson, 1994) and persons remember stimuli that arouse strong emotions better, while non-emotive stimuli tend to be easily forgotten (Bower, How might emotions affect learning, 1992).

The next process is behavioural learning (Jansson-Boyd, 2010); this is focused on observable behaviour instead of the mental processes commonly explored in classical psychology models. Behavioural learning analysis is therefore widely employed in analysing consumer behaviour, with classic and operant conditioning being its most common methods. This area of psychology have been commonly decried as focused in controlling the consumer actions to the benefit of the seller. However, it can provide a robust understanding of the processes and mediums necessary to transmit a message to the public and has been used by different government to achieve specific behavioural objectives in family planning, disease prevention, environment protection and better treatment of minorities (Thomson O. , 1999).

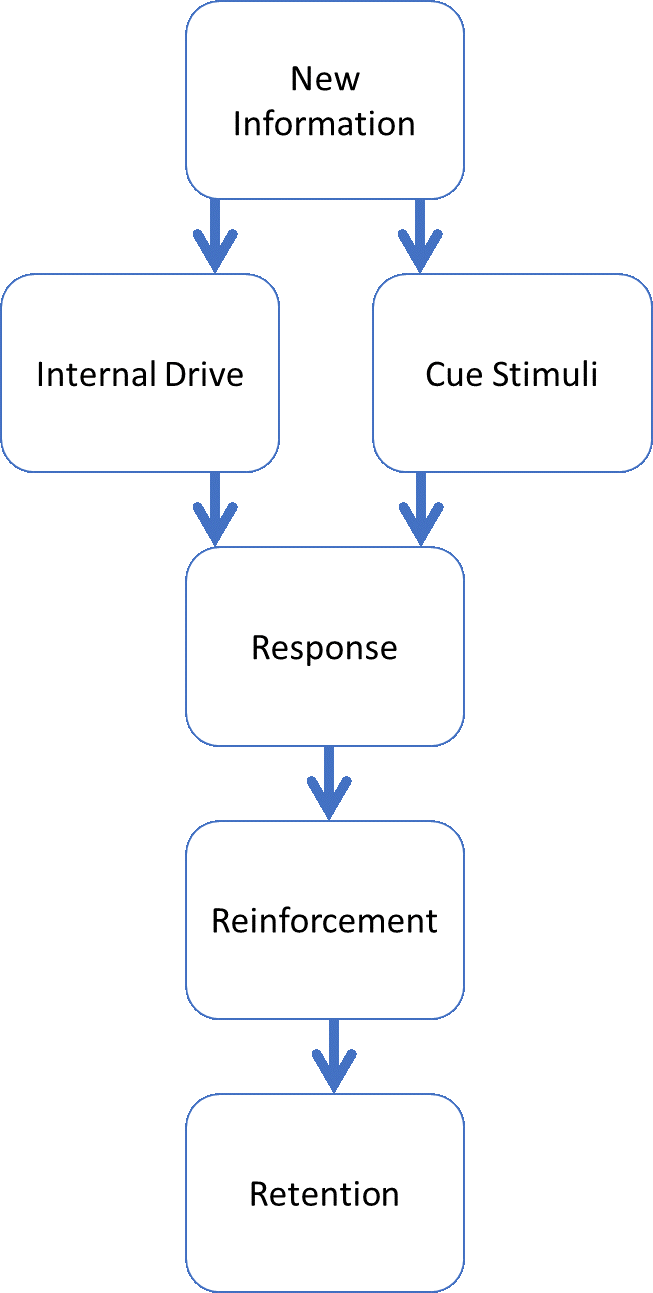


Figure The elements of learning (Jansson-Boyd, 2010)

As can be seen in Figure 5, the model for the learning process starts with the presentation of new information about the product. Then the internal drive (guided by emotions) and external stimuli create a response in the consumer. The result of the consumer’s decision then must be reinforced by a positive experience to ensure the material is learned and stored in long-term memory. The idea that emotion is an intrinsic part of how we learn has been evaluated since Darwin’s theory (Darwin, The expression of the emotions in man and animals, New York: D, 1872) and in the marketing area has been corroborated by several studies like (Bolls, Muehling, & Yoon, 2003), (Eighmey & Sar, 2007), (Schulz, et al., 2004) that suggest sentiments, emotions and sympathy made consumers more open to suggestions. However, not all studies are focused on selling products, several times the government itself has employed this models to achieve specific behavioural objectives (Peterson, 1939), (O'Donnell & Jowett, 1986), (Thomson O. , 1999), (Hilton, 2007). These objectives can include moral propaganda like family planning, disease prevention, environment protection and better treatment for minorities (Thomson O. , 1999). Finally, these ideas can be summarized in Rousseau’s words “*Whoever makes it his business to give laws to a people must know how to sway opinion and govern the passions of men*” (Thomson O. , 1999).

When applying these ideas to the design of PILs, the need to design documents that are clear and interesting, catch the reader’s attention, and help the reader to take action becomes apparent (The Writer), (PULSE). This contrasts with the current recommendations for writing PILs provided by medical guidelines (MRC, 2016), which seek uniformity, objective and respectful or formal texts.

## Clustering techniques for identifying core characteristics

As previously mentioned, clustering relates to the idea of grouping objects in accordance with a similarity metric. This methodology was first implemented not in the Web but in the anthropology area to evaluate the quantitative expression of cultural relationships (Clements, 1954) and has been adopted by practically all areas of research as a response to the need of processing huge amounts of data. Formally defined, clustering can be expressed as:

Definition-1. Given a set of elements S = e1, e2, ..., em , a subset C is called a cluster if ∀x, y∈C, ∀z∉C→d(x, z) < d(x, y), where d:S×S→R is a distance function between the elements in S.

From the previous definition, it can be observed that the formation of the clusters depends on the definition of the similarity metric, thus making clustering more of a task to be solved than a particular method or process. This has led to numerous similarity measures by researchers seeking to adapt the analysis to the particulars of their data. These general approaches to define a cluster can be categorized into (Xiong, Wu, & Chen, 2009):

* Evaluating connectivity distances
* Making a graph interpretation
* Employing statistical distributions
* Analysing density regions
* Structuring based in members and attributes
* Employing mean vectors

In addition, the methodologies can be categorized based on how rigorously they apply cluster membership (Sarle, 1990). From the most rigorous, where all elements must belong to a single cluster, to allowing elements not to belong in any cluster, to also belong to parent cluster and finally to have no restrictions on the number or type of cluster an element can belong to.

In this research, the words in the PILs are defined as the elements of analysis while the degree of similarity is given by the number of words two documents share. By seeking the relationships between the appearance of certain words or combinations of words with readability metrics (attention, recall and understanding) we aim to provide the Principal Investigators with valuable insights when making their PILs.

# Analysis of the characteristics of PILs text

## Research Questions

1. Is there any emotive difference between PILs and other texts that make them harder to understand?
2. Are there any characteristics with significant associations to low readability and poor recruitment?
3. Is it feasible to employ Web techniques to revise PILs text?

## Aims

1. To identify the ratio of words related to emotion or sentiment present in the PILs and compare them with other kinds of text expected to be understood by public audiences.
2. To identify the differences between PILs and other texts that could hinder the reader capacity to understand the essential aspects of the trial.

## Justification

Classical and applied psychological studies have associated the process of learning, attention and recall (memory) to emotional stimuli since Darwin’s theory on emotion (Darwin, The expression of the emotions in man and animals, New York: D, 1872), (Starch, 1914), (Ekman P. , 1973), (Kreutzer & Charlesworth, 1973), (Averill, 1980), (Blaney, 1986), (Ekman & Davidson, 1994), (Eysenck, 2012), (Adams, 1916), (Jansson-Boyd, 2010), (Escudero-Carretero, et al., 2013), (Eighmey & Sar, 2007). Based on these studies, the idea of creating objective, non-biased and respectful PILs by excluding emotive words from the text could significantly hinder the document’s capacity to retain the attention of the reader (Jansson-Boyd, 2010) and the patient’s ability to recall the essential aspects of the trial (Bower, How might emotions affect learning, 1992) needed to make a meaningful decision.

The inherent purpose of news articles is to present information to the public in a manner that is easy to understand. Journalism -just like research- is often based on information gathering and problem solving, processes which follow a similar structure by answering the Six-W questions (Hart, 2002): who, what, where, when, why and how? Moreover, the articles’ capacity to transmit information can be perceived by their social impact (Dixon & Linz, 2002), (Entman & Rojecki, 2001). Thus, the objective of this study is to compare the number of words related to each of the primary emotions (Plutchik, 1984) present in PILs and in news articles from the BBC, Hello Magazine and Daily Mail.

## Background & Rationale

Patient information PILs are recognised as an essential part of clinical trials in the UK (NHS, 2017). They should help to “ensure that all those who are invited to take part in a research study have been adequately informed” (NHS, 2017). On the other hand, most research in the last decade has found the PILs to lack objective assessments of quality (Moult, Franck, & Brady, 2004), have poor readability (Nicholls, Hankins, Hooley, & Smith, 2009), (Escudero-Carretero, et al., 2013) and are of insufficient quality in all categories other than covering the legal and ethical requirements (Reinert, et al., 2014).

A common perception by recruiters, nurses, doctors, researchers and ethics committee members is that the PILs have no actual impact on the patient decision and are commonly not read or remembered (Poplas-Susíc, Klemenc-Ketis, Kersnik, & others, 2014). Even when the current guidance encourages the production of higher-quality PILs by employing Public Involvement (PPI) groups in their design (NHS, 2017) and several studies have sought to improve the PIL quality (Reid, et al., 1995), (Knapp, Raynor, Silcock, & Parkinson, 2011), (Escudero-Carretero, et al., 2013), (Gillies, Huang, Skea, Brehaut, & Cotton, 2014), (Reinert, et al., 2014), these issues still remain valid and have become priority topics of research TMRN (Healy, et al., 2018).

The studies on the area have focused on evaluating the PIL quality through quantitative readability metrics like length, font-size and sentence structure (Knapp, Raynor, Silcock, & Parkinson, 2011), (Reinert, et al., 2014) that show moderate success but still fail to create adequate PILs. This study adds the quantification of terms related to emotions as psychological literature has related the process of cognition and memory to emotional stimuli since Darwin theory of emotion (Darwin, The expression of the emotions in man and animals, New York: D, 1872), (Starch, 1914), (Ekman P. , 1973), (Kreutzer & Charlesworth, 1973), (Averill, 1980), (Blaney, 1986), (Ekman & Davidson, 1994), (Eysenck, 2012).

## Methods

This is a cross-sectional observational study of the Patient Information PILs (PILs) from Randomised Controlled Trials (RCTs) supported by the HRA between 2000-2014 with publicly available PILs. Information on the trials was provided by NETSCC.

### Sample & Recruitment

A set of 58 publicly available Patient Information Leaflets and the details for their Randomized Controlled Trials supported by the HRA between 2000-2004 was collected for analysis with help of the NETSCC.

### Limitations

To be added

### Procedure

* Information on 181 trials was provided by NETSCC.
* 72 trial PILs with publicly available PILS were selected.
* PILs from trials with a randomisation process that was not focused on the patient (i.e. cluster and cohort randomised trials) were discarded.
* Where multiple PILs were provided for the same trial, only the PIL which targeted the decision maker was used.
* PDF documents with the text of 58 PILs were created by splitting the PIL section of the trial protocol when necessary.
* The text of each PIL was extracted from the PDF PIL by employing the Online OCR platform (OnlineOCR, 2016) into a simple text file with UTF-8 encoding.
* A visual inspection of the PIL text was done:
  + Correct sentence structure when the PIL employed multiple columns (3 cases)
  + Delete extraneous text which was not part of the original PIL (Headers and footers added when the PIL was inserted into the protocol file)
* A Python program was designed to clear numbers and symbols from the file
* A weighted list of words present in each PIL was made using a Python program
* 60 news articles were saved offline by a Python spider program which crawled the homepages from the BBC, Daily News and Hello Magazine sites and stored the main articles between 06/04/17 and 21/04/17.
* The text of each news article was cleaned numbers and symbols with a Python program
* A weighted list of the terms that appeared in each news article was created
* The Canada National Research Centre (NRC) Emotion Lexicon (EmoLex) (Mohammad & Turney, NRC emotion lexicon, 2013) was used to quantify the number of emotive terms present in the PILs and news articles.
* The number of terms associated with each principal emotion present in the PIL and news articles was compared.
* Linear regression models were used to evaluate the association of each emotion with the trial recruitment rates.
* Clustering analysis was used to identify PIL words associated with high or low recruitment rates.
* Linear regression was used to determine the significance of the previously identified words.

### Measures & Analysis

#### Making a quantitative evaluation of the NRC Emotion Lexicon validity:

There were concerns that a general emotion lexicon would not prove to be a valid tool in the analysis of PILs because the use of passive voice may change the sentence wording and may inhibit the recognition of emotional words. A verification test was carried out to check the number of recognized words was similar in the PILs and news articles.

#### Employing the NRC emotion lexicon to analyse the number of emotion-related words present in the PILs:

A selection of 60 articles each from BBC, Daily Mail and Hello Magazine were acquired from their respective webpages. The emotion displayed by these articles was analysed with the NRC Lexicon and their results compared with the previous results for the PILs obtained during my MSc research. In the PhD, a study with commonly read texts was deemed necessary to have a better emotional comparison with the PILs.

The previous study was part of my MSc research where employed the NRC Emotion Lexicon version 0.92 (Mohammad & Turney, Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon, 2010) to explore 58 PILs corresponding to patient randomised RCT trials funded by the HTA between 2000 and 2014. The study quantified the proportion of words related to each of Plutchik’s emotion categories and the positive-negative sentiments. The results included a comparison with emotive texts in which the NRC Emotion Lexicon has been used previously by the lexicon’s creator research group. However, the texts analysed by the NRC research group were deemed too emotional to be appropriate comparisons for the PILs.

#### Employing K-means clustering to find a set of PIL-words that correlate with recruitment rates:

Currently, this study has only been carried once during my MSc research (Annex 1). The text in the PILs was separated into single words and the words were clustered by their ratio of appearance in PILs with good/poor recruitment (where poor recruitment is defined by the failure to recruit at least 80% of the original samples) by employing the K-means algorithm.

#### Employing a linear regression model to refine the word set into a list that presents significant correlation with recruitment rates:

Employing a linear regression model was considered necessary to further refine the results obtained by cluster analysis. Once the clusters of phrases (combination of words) from different RCT PILs that present possible relations with recruitment, attention, recall and understanding have been identified with clustering analysis, a linear regression model using the stepwise method will be employed to ascertain the level of statistical significance of their relationships. A similar study was employed during my MSc research to identify a list of 23 PIL-words that presented a statistically significant relationship with recruitment in RCTs (Annex 2).

## Results

The first step on my PhD research was to check the validity of employing a general lexicon on PILs. During my previous MSc research, it was observed that the use of passive voice could change phrase wording. Therefore, it was necessary to ascertain that a similar number of words were recognized in the PILs as in the news articles. Thus, a selection of 60 articles from each BBC, Daily Mail, and Hello Magazine websites were acquired. The following figures give a general description of the characteristics of those texts:

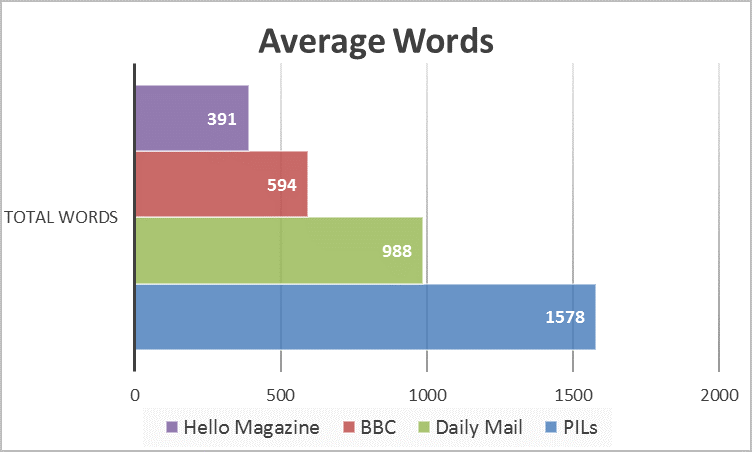


Figure Average number of words in news articles and PILs

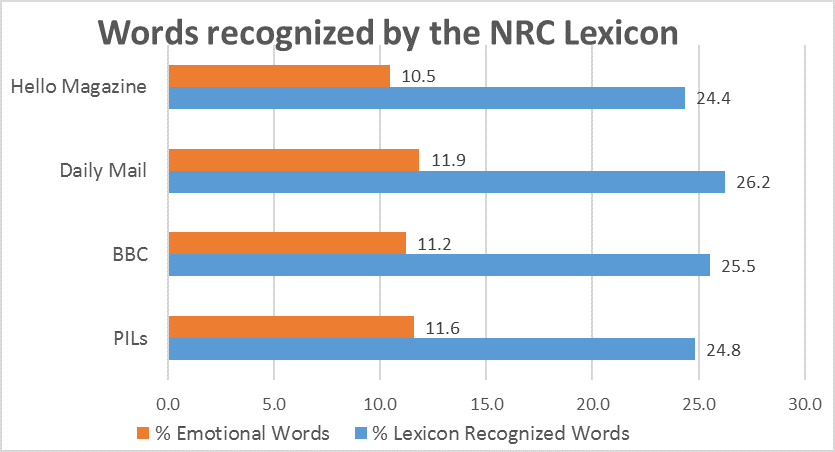


Figure Average percentage of lexicon words identified as emotional

While the length of the PILs was significantly larger than the length of the news articles Fig 7, a similar percentage of words were identified by the lexicon in all documents Fig 8, with a similar ratio of them being considered emotional (words containing an association to at least one emotion category). Thus, the NRC Emotion Lexicon was deemed to be a valid instrument to quantify the emotion present in the PIL text.

Next, an emotion and polarity analysis of the content of the texts was carried out. This was done by quantifying the ratio of text related to each of the central emotion categories on Plutchik’s Wheel of Emotion (Plutchik, 1984) and to the positive-negative sentiment categories. This resulted on a comparison of the level of emotion and sentiment-related words present in the PILs with text the patient could be expected to read in their everyday life. A general overview of the results can be seen in the following table:

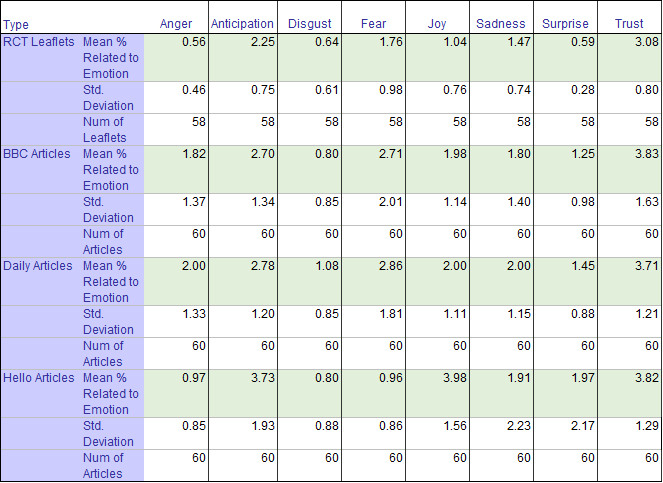


Figure Comparison of PILs with news articles from 3 sources

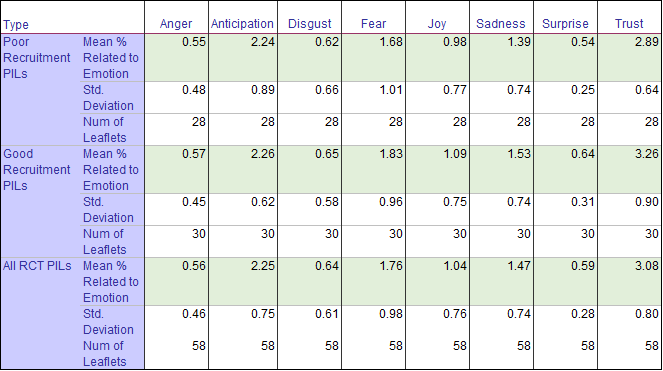


Figure Comparison of PILs recruiting well (at least 80% of planned sample) and badly

The results of this study reflect interesting results when paired with the results achieved in the previous MSc research. The general comparison of news articles with my previous MSc results show that there is a lower variation in the quantity of words related to each emotion category in the PILs than in other texts (Figure 6,Figure 7).

In addition, there is a lower number of words related to the middle layer categories of Plutchiks’ Wheel of Emotion, illustrated in Figure 1. However, this was compensated by being the texts with the most positive-sentiment words (Figure 9). The visualisation of the analysis results was divided in two sections, a comparison of Plutchiks’ emotion categories and one for the positive-negative sentiments. For the first, radial charts were employed to visualize the quantity of words related to the emotion categories while a clustered bar model was deemed more appropriate for the polarity analysis:

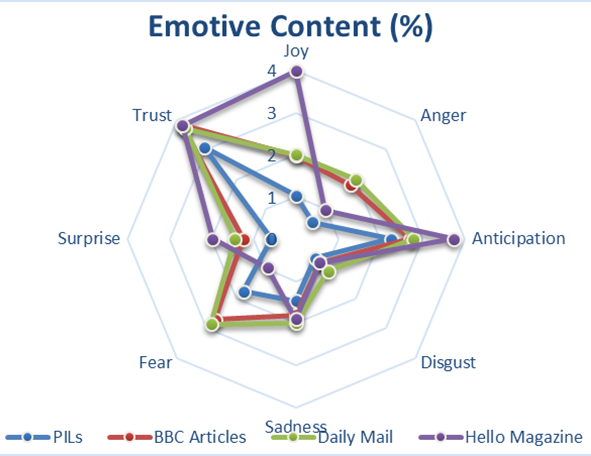


Figure Percentage of words related to each emotion category in the news articles & PILs

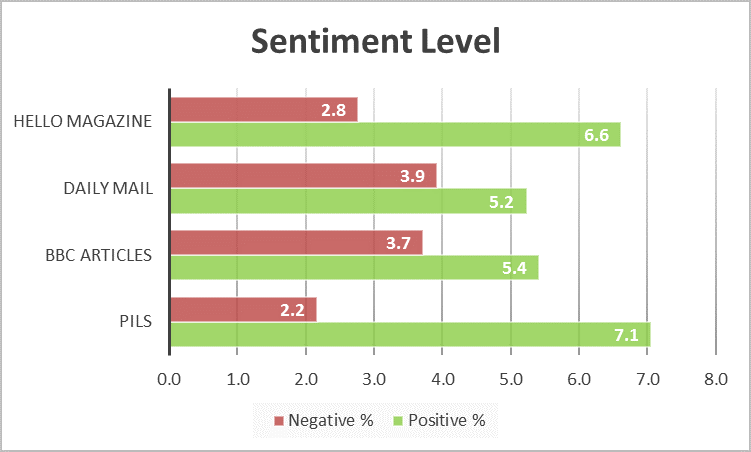


Figure Percentage of words related to Positive and Negative sentiments in the control documents & PILs

When compared to the results found in the previous MSc. research (Figure 10), a clear reduction in the difference between the ratio of emotion present in the texts can be appreciated.

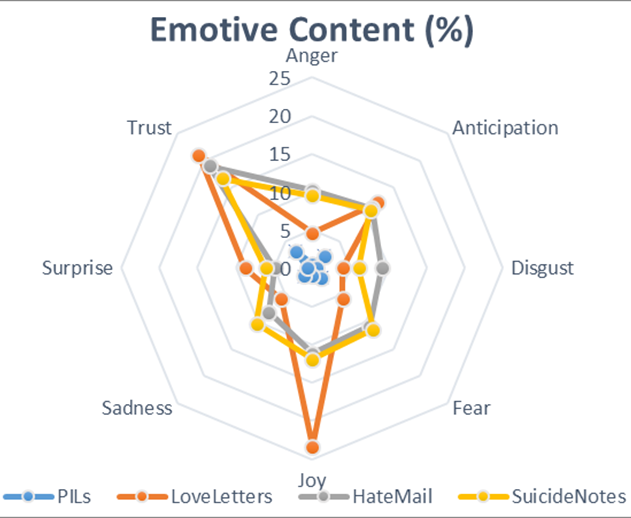


Figure Percentage of words related to each emotion category in love letters, hate mail, suicide notes & PILs (MSc research)

The previous MSc results were obtained by comparing the PILs with published results of studies employing the NRC lexicon to analyse emotive texts (love letters, hate mail and suicide notes) (Mohammad & Yang, Tracking sentiment in mail: how genders differ on emotional axes, 2011). These resources were deemed too emotional to be appropriate comparisons to PILs. However, the current analysis still shows significant lower quantities of emotional content other than positive sentiment in the PILs and lower variations in the words related to the emotion categories when compared with common texts.

## Discussion

This study compared the emotive content in the PILs with the emotional spectrum found in love letters and hate mail reported by independent researchers. The proportion of words related to emotional categories was found to be extremely low in comparison. In addition, a comparison of the emotive content in the PILs with news articles from Daily Mail, BBC and Hello Magazine was made. News articles’ emotive content varied depending on the news category and the intended audience, in contrast the PILs generally present lower variations in the emotional content. In addition, the PILs have a higher positive sentiment than the articles, but are lacking on all other emotional categories. Finally, the capability of the lexicon to recognize words in the PILs was assessed, as there was the suspicion that formal language may hinder it as the words’ spelling changed. The results show that a similar number of words is recognized in the PILs as in the news articles (around 25%).

The lack of emotional variation present in the PIL text makes a clear difference with other text sources that the patient may encounter in their everyday life, which employ different ratios of emotive text depending on the article and the audience. When the PILs were visually observed, it was found this occurred because the PILs employed a highly formal language and in many cases templates for the PILs.

The literature suggests that documents which lack a certain level of emotion are harder to understand and remember (Jansson-Boyd, 2010), (DeWall & Myers), (The Writer), (PULSE), that the commonly employed templates and language produce documents that require high levels of education to read/understand and that the length of the text is excessive when trying to maintain the readers’ attention. All the previous points bring into question the role of the PILs as a support to help patients make a decision.

### Strengths & Weaknesses

To be Added

# Analysing Public Comments on PILs with Low Readability and Poor Recruitment Rates.

## Summary

To be added

## Research Questions

1. What type of comments do public reviewers give on low readability PILs from poor recruiting trials?
2. Is it possible to determine reader understanding of the trial features from their comments and subjective perception of quality?
3. If so, how many reviewers would be necessary to cover at least 80% of the comments topics significantly associated with reviewer understanding or quality perception?

## Aims

1. To determine the type and themes of comments given by public reviewers on PILs from RCTs with low readability or poor recruitment
2. To collect PPI comments into a database or future use in pre-reviewing PILs for RCTs
3. To find possible correlations between PPI comments, quantitative measures of content analysis and the PIL’s capacity to inform on essential aspects of the trial.
4. To measure the level of coverage of small groups of public reviewers (3-9 people) have when commenting on trial PILs.

## Background & Rationale

Clinical trials have become a cornerstone for maintaining high quality in the health-care systems of developed countries (Lovato, Hill, Hertert, Hunninghake, & Probstfield, 1997). They enable researchers to compare the effects of new drugs and treatments against those that are currently employed, to improve the health-care of the general population by developing new guidelines and practices (NHS, 2017). On the other hand, their very nature implies a risk for the patients that choose to participate in them of either receiving a sub-optimal treatment or suffering previously undiscovered side-effects (Moore & Savage, 2002). Thus, it is of great importance the creation of regulations to ensure the patient is aware of the risks and enforce the ethical practice during recruitment (MRC, 2016).

One of the core elements in informing the patient is the patient information PIL (PIL) (Knapp, Raynor, Silcock, & Parkinson, 2011) with all clinical research studies that involve patients in the UK being asked to develop one. The Health Research Authority (HRA) defines the role of PILs as “should support the consent process by helping to ensure that all those who are invited to take part in a research study have been adequately informed” (HRA, 2017) and it states that effective informing enables the participant to make appropriate decisions and that the information should be understandable by the participant and consider their own circumstances (HRA, 2017), (MHRA, 2016).

These PILs are usually designed by following the guidance provided by the HRA, are commonly based on templates (HRA, 2017) and are reviewed by an ethics panel as part of the research procedure in the UK. To further ensure a high quality the researchers are recommended to employ Public and Patient Involvement (PPI) groups to review the PILs (HRA, 2017).

However, there is no formal requirement to do so, the public reviewers are commonly below 9 people with no formal assessment of the effect of their comments on PIL quality and the recommended £25 per hour per person pay (National Institute for Health Research -NHS, 2014), (School for Primary Care Research, 2017) can be excessive for pragmatic trials. Even more, several research studies have found that we do not have a standardized method for assessing the quality of PILs (Moult, Franck, & Brady, 2004) and that current PILs do not communicate the information effectively (Gillies, Huang, Skea, Brehaut, & Cotton, 2014), (Reinert, et al., 2014), (Knapp, Raynor, Silcock, & Parkinson, 2011). While RCTs’ PILs are considered essential for valid consent (Knapp, Raynor, Silcock, & Parkinson, 2011), in general, they lack the capability to meaningfully inform the patient (Gillies, Huang, Skea, Brehaut, & Cotton, 2014) and in some cases, they may not enable valid consent (Knapp, Raynor, Silcock, & Parkinson, 2011). In particular, the formal tone of the PILs was found to hinder the readability of the text (Reinert, et al., 2014).

The Health Research Board Trials Methodology and Networks (TMRN) working with the James Lind Alliance and the TrialBank convened the Priority group to identify several priority questions including quantifying the effects that public reviewers have on PILs and RCTs (Healy, et al., 2018). Specifically, identifying which information should be communicated, assessing the effect of PPI collaboration on recruitment rates and finding the best methods to deliver information are among the top five questions identified by this JLA priority-setting panel (Healy, et al., 2018). Therefore, it is our intention to build a tool that supports public reviewers reviewing PILs for RCTs in the UK answer these questions.

## Methods

This is a cross-sectional observational study of the Patient Information PILs (PILs) from Randomised Controlled Trials (RCTs) and their capacity to inform people about essential aspects of the trial.

### Sample

The PILs will be analysed by 20 public participants as it has been established by talking with PPI officers in the Faculty of Medicine and the Wellcome Trust Clinic that it is common to have public reviewers of less than 10 people but no formal assessment of the number of reviewers needed has been done. We anticipate that 20 participants will be sufficient to reach saturation on the comments they generate.

### Recruitment

24 participants were recruited to review and comment on 4 PILs from RCTs which failed to recruit at least 15% of their original planned sample.

Collected the following demographic participant information:

1. Previous participation in RCTs [Nominal]
2. Interest in participating in a RCT [Nominal]
3. Age group [Ordinal]
4. Education level [Ordinal]
5. Health level [Ordinal]
6. GP visits in the last year [Ordinal]
7. Gender [Nominal]
8. Origin [Nominal]

Having the participants complete the following tasks:

1. Read the information about the study and fill the joining form.
2. Review the PILs
   1. Read the PIL text
   2. Highlight sections of text were issues are found
   3. Comment on the highlighted sections of text
   4. Describe the general quality of the PIL
      1. Grade the overall PIL quality on a unipolar scale (0-10)
      2. Provide a general description on the PIL issues
3. Answer a questionnaire on the essential information present in the PIL based on topics identified by the Expanded EQIP scale (Charvet-Berard, Chopard, & Perneger, 2008)as relevant to measuring PIL quality.

### Procedure

We have selected a set of four PILs from RCTs supported by the NIHR HTA programme between 2006 and 2009 that failed to recruit at least 15% of their initial planned samples. This selection will permit identification of whether there are common issues in PILs from trial interventions with extremely poor recruitment that could be recognised by analysing the public comments on their PILs. The selected PILs are from interventional phase IV trials that compare treatments that were currently in use and had similar expected populations. These PILs will be reviewed by public participants using the following process:

1. Contact public participants: around 20 participants will be recruited by posting ads in social platforms, e.g. Facebook & Tweeter, establishing a group in the Millennium Third Age Centre (3AC), and potentially inviting the ECS and FoM communities to participate by sticking ads in the community blackboards. In case of lower recruitment than estimated Gary Hickey from the INVOLVE group will contact public members on our behalf.
2. The participants will be asked to provide comments on each PIL’s text; each comment will be added to a database (the CARPI -Computer Assisted Reviewer for Public Involvement- database) and employed in the 2nd phase of the project.
3. The participants will give an overall quality grade for each PIL. This grade will be an estimate given by the participant through a unipolar analogue scale with a range of 0-10.
4. The PIL’s capacity to inform will be assessed by having the participants take a quiz based on the Knapp scale (Knapp, Raynor, Silcock, & Parkinson, 2011) via an online survey.
5. The comments will be thematically analysed by employing Applied Thematic Analysis (ATA) and following the coding methodology provided by Saldaña (Saldaña, 2015), which include methodologies to analyse qualitative data in a systematic and rigorous way.
6. The comments will be stored anonymously in the CARPI database and used in the next phase of the project to identify sections of PILs for trials currently under review by a PPI group that are similar to sections previously commented on by our participants.

No specific selection criteria will be used for the participants, only that they are adult residents of the UK. The inclusion of all segments of the population should help identify differences in how the text is perceived. In particular, the understanding of PILs by populations with lower-levels of academic background or with English as a second language are hypothesized to be hindered more by the formal language of the PILs.

The PILs’ text will be provided to the participants as Word/Open Office/PDF documents that have been standardized by:

* Removing all images
* Removing all names or trial identifiers
* Employing Times New Roman 12 as the default font
* 1.5-line spacing

Each participant will be asked to provide their comments by selecting a section of the text and making a comment/note in the Word/Open Office menu displayed when right-clicking the area. In case this method is not possible, the participant will have the option to print the text, mark the text and write a number corresponding to their comment in the area above the section, and then write their comment on a separate comment page provided with the document. Should this occur, the participant will be asked to send a scan or photo of all the pages that have at least one comment from the PIL text and the comment pages to the researcher’s institutional e-mail account.

The participants will then be asked to provide an overall quality grade for each document and to answer an online survey using EQIP to evaluate their understanding of the information in each document. The participants will need to write their user id, given at the end of the registration process, and the overall grade on the first page of the document.

All data will be stored anonymously in the University of Southampton institutional account of the researcher. No personal information will be asked of the participant. The information will be used as aggregates to identify correlations with several quantitative text analysis metrics. Also, the participants’ comments will be used to give suggestions when similar sections are identified in other PILs. The final reports of this phase were submitted as a PhD paper to the HEALTAC’18 conference in Manchester UK.

The comments acquired from the participants will be analysed in an iterative process by identifying the comment features (themes) and defining their boundaries (text segmentation). These themes are classified in two categories:

* Structural themes: that are imposed by the research design, in our case, the comments can be structurally classified as:
  + Suggestion to replace part of a sentence
  + Instruction to clarify a sentence
  + Qualitative assessment of a sentence
* Emergent themes: are the particular topics that will emerge when observing the comments in the context determined by the research design. I.e. they can be seen as the justification for each comment in our research.

Once the themes for each comment have been identified the similarity of the comments will be assessed to determine the optimal number of reviewers to cover all the topics. This similarity analysis will be done by employing K-means clustering, as this technique has been found to be useful in revealing groups of participants with similar profiles when employed on coded qualitative data with quantitative and qualitative features (Henry, Dymnicki, Mohatt, Allen, & Kelly, 2015).

### Analysis & Metrics

#### Analysis of the comments given by the reviewers.

In a first step we classified the comments into two categories, comments that were given on specific sections of text and those that referred to the overall quality of the document. Then, we employed qualitative coding to analyse the type and structure of the comments provided by the reviewers. Specifically, we employed elemental (in-vivo), affective (emotion) and language (motif) coding methods. In-vivo coding (Charmaz, 2014) was used initially to recognize the overall categories of the comments by assessing the similarity of the phrasing of the comments. These classes were refined by employing motif (Mello, 2002) and emotion (Prus, 1996) coding. Emotion coding permitted the classification of comments by labelling which emotions were perceived in the texts, while motif coding added the recognition of patterns on the subject of the comments. In the initial step the comments and the PIL sections associated with them were taken into account:

Text in the PIL: “From previous studies, we don’t expect there to be any difference in effectiveness between these two (treatments)”

Comment: “why are they doing it then?”

Then, the emotion the comment expressed is added:

Emotion: Distrust, annoyance

Finally, motif coding was employed to ascertain the overall topic and motive of the comments. In literary works motifs are seen as recurrent ideas, symbols or actions that develops or explain a theme (Mello, 2002), (Thomson S. , 1997) (Saldaña, 2015). A motif is defined as the “smallest element in a tale that has meaning on any other tale” (Saldaña, 2015):

Topic: Study risk and disadvantages  
Motif: Cogitation (induce reflection on the investigator)

These coding strategies were used to classify 159 specific comments. These were found to be associated with the percentage of correct answers (Expanded EQIP scale) obtained by the reviewers. When a comment was found to include more than one topic or motif, multiple instances of the comment were generated to account for each one:

**Comment**: “People will need to know if there are risks or disadvantages not -we don’t think there are- this is an open statement”

Table Comment with multiple motifs

|  |  |
| --- | --- |
| 1st Instance | 2nd Instance |
| Text: “We don’t think there are any risks or disadvantages”  Instance comment: “People will need to know if there are risks or disadvantages not -we don’t think there are-”  Emotion: Distrust, anger  Topic: Study risks and disadvantages  Motif 1: Contention (confront the researcher ideas and views) | **Text**: “We don’t think there are any risks or disadvantages”  **Instance comment**: “this is an open statement”  **Emotion**: Distrust, anger  **Topic**: Study risks and disadvantages  **Motif 2**: Change wording (do not use open statements) |

In this example a comment generated two different instances of motifs present in the comment, the idea that the study did not carry any risks for the participants and that open statements should be avoided.

#### Employ statistical analysis to determine the number of public reviewers needed to ensure a good coverage of the commented topics

Taking into account only the set of comments associated with correct EQIP answers (n=159), we determined the probability that by taking an 8 reviewers we could achieve a coverage of at least 80% of the comments. First, we computed the percentage of specific comments each reviewer provided, then the minimum set of cases that covered 80% of the comments was obtained (X=6). The probability of this set occurring was determined by assessing the ratio of permutations in which the event occurs against the total number of permutations.

#### Employ linear regression model to identify factors significantly associated with the reviewers’ perception of quality & their understanding of the trial’s essential features.

The comparison of demographic factors and the type of comments with the perception of PIL quality and participant understanding was done by employing linear regression models. The perception of quality was subjectively assessed by the participant via a unipolar scale (0-10). The understanding of the trial features was evaluated by having the participants answer a questionnaire derived from the topics found as relevant to PIL quality by the validated Expanded EQIP scale (Charvet-Berard, Chopard, & Perneger, 2008).

## Results

### Sample detail

A total of 24 public reviewers were recruited by employing the social media platforms of the 3rd Millennium Age Centre who were classified into three age groups:

Table Participants’ Age Groups

|  |  |  |
| --- | --- | --- |
|  | No. of Participants | Percent |
| 18-30 | 6 | 26 % |
| 31-44 | 8 | 35 % |
| 45+ | 6 | 26 % |
| Valid | 20 | 86 % |
| Missing | 4 | 14 % |
| Total | 24 | 100 % |

From the recruited participants 52% were from the UK, 9% European and 26% International. 48% identified themselves as female, 26% as male and 9% not stated. The recruited participants had a wide range of education levels, with similar numbers of participants in each category:

Table Participants' Education Level

|  |  |  |
| --- | --- | --- |
|  | No. of Participants | Percent |
| No GCSEs | 4 | 17 % |
| GCSE / O-levels | 4 | 17 % |
| A-levels | 3 | 13 % |
| Undergrad | 2 | 9 % |
| Grad | 5 | 22 % |
| Postgrad | 2 | 9 % |
| Valid | 20 | 86 % |
| Missing | 4 | 14 % |
| Total | 24 | 100 % |

The comments from participants who refused to provide demographic information were excluded from the regression analysis measuring the association of participant education level and age group with their understanding and PIL quality perception.

### Comment analysis: How many reviewers are needed to obtain a good coverage of comments?

Analysis of the data shows that reviewers gave two different kinds of comments: general comments on the overall structure of the PILs and comments on specific sections of the PILs. The reviewers gave on average 4 general comments (1 comment per PIL) with a standard deviation of 2, only 10 reviewers gave comments on specific section of the PILs with an average of 16 specific comments per participant and a standard deviation of 10 (Figure 11). However, 49.6% of these specific comments were given to the first PIL, with sharp reductions in subsequent PILs (28.3%, 15.7%, 6.3%).

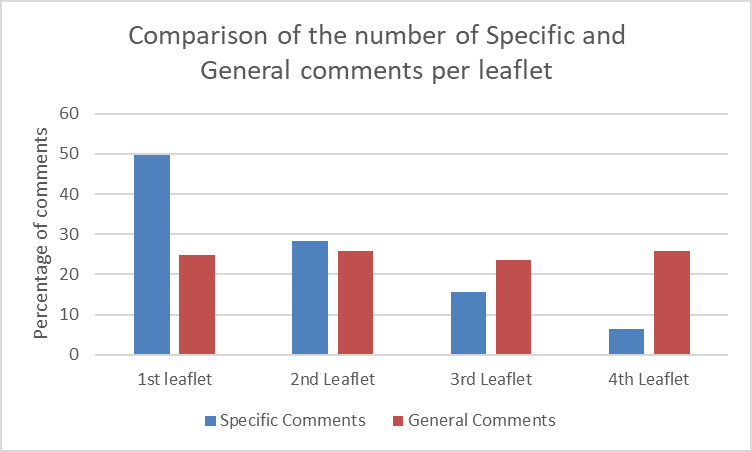


Figure Comparison of the number of general and specific comments given to each PIL

From the total of 240 comments captured, only those associated with specific sections of the PILs (n=159) were found to be significantly associated with percentage of correct answers and the quality grades given by the reviewers Figure 12. Selecting this set of 159 specific comments as our base, the percentage contribution of each reviewer was computed:

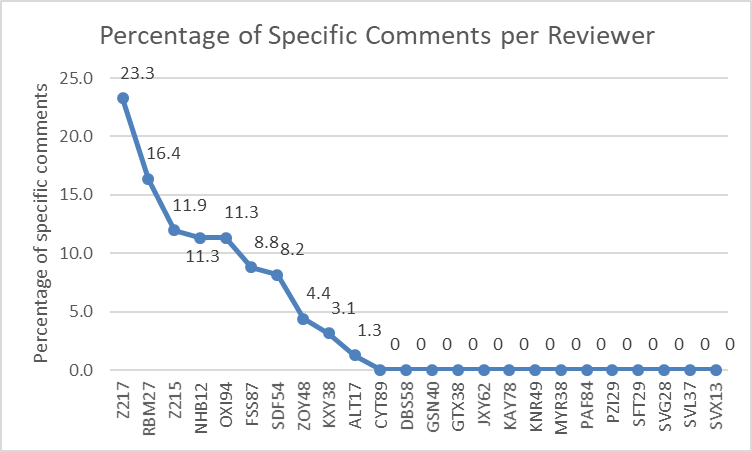


Figure Contribution of specific comments per reviewer

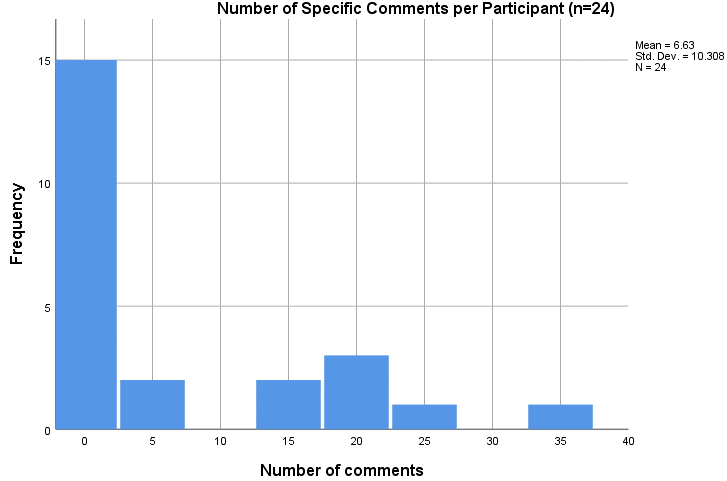


Figure 13 Distribution of specific comments taking the full sample of reviewers

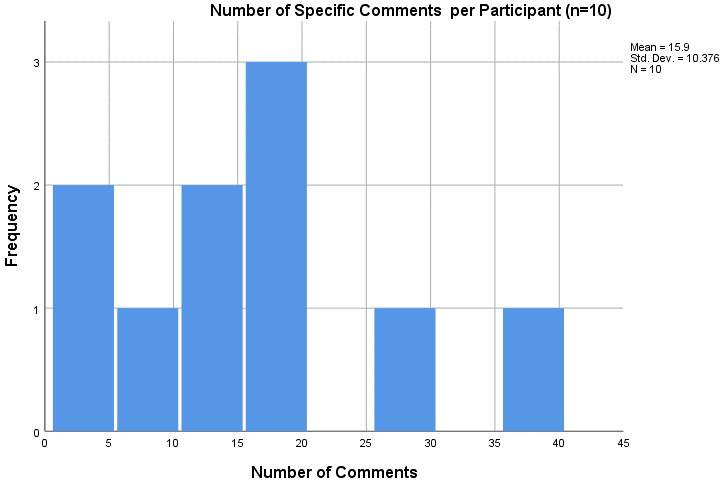


Figure Distribution of comments when excluding reviewers who did not gave any specific comment

As it can be appreciated in the histograms the number of specific comments given by the participants did not follow a normal distribution Figure 13 even when only accounting for the data of those reviewers that provided any (n=10) Figure 14. The median number of specific comments per participant was 16 with an interquartile range of 14, a standard deviation of 10.4 and mean of 15.9.

Using the descriptive statistics for mean and standard deviation of number of comments on specific sections, we calculated the margin of error for the contributions (percentage of specific comments) of the reviewers given an 80% confidence interval (z-value = 1.28):

*MOE* = *z*\*(σ/)/(√(n)) = 1.28\*(6.5/)/(√(10)) = 2.6

With this information, we employed set theory (Bollobás, 1998), (Bondy & Murty, 1976) to model the information. Where R represent the set of reviewers, S contains the sets of comments on specific sections given by each reviewer and G is defined by the sets of overall comments given by each reviewer. The set of all comments given by each reviewer can be defined as *C* = {*Ci* ∈ *C* | *Si* ∈ *S*,  *Gi* ∈ *G* *Gi* = *Si*∪*Gi*} where *Si* *and* *Gi* represent the set of comments given by the reviewer *i* on the specific sections and the overall document respectively. We define the set of reviewers who gave at least one comment on an specific section, RS:

*RS* = {*r* | *r* ∈ *R*,  |*Sr*| > 0}

Then we compute the contribution of each reviewer r to the total number of specific comments with the following formula:

*fsc*(*r*, *S*) = (|Sr|)/(∑*i*= 0|*S*||*Si*|)

And we employ this information to obtain a minimal set of reviewers, *MSR*, capable of covering at least 80% of the specific comments:

*MSR* = *min*(*A* ⊆ *RS* | ∑*i*= 0|*A*|*fsc*(*ai*, *S*) > 0.8,  *where* *a* ∈ *A*)

This set was found to be not unique, with four possible solutions composed of 6 reviewers each (|*MSR*| = 6) which accounted for at least 79.9% of the comments associated with specific sections. The probability of these sets occurring was calculated by obtaining the ratio of possible solutions over the total number of unique combinations:

*P*(*MSR*) = (numSol)/(10*C*6) = (4/)/(210) = 0.02

This shows a 98% probability of not covering at least 80% of the relevant comments with reviewer groups composed of less than 7 members. A group of 7 people would have a coverage of at least 46% and a maximum coverage of 91% of the specific comments in our study. To obtain the average percentage of specific comments covered by a group of 7 reviewers we computed all the unique possible combinations in which such a set could be formed, following the unique combination formula:

*nCm* = (n!/)/((*n* − *m*)!*m*!) = (10!/)/(3!\*7!) = 120

Once the groups were formed the percentage of comments covered by each group was obtained. The subgroups formed in this way covered on average 70% of the specific comments with a standard deviation of 9.5. The probability of covering at least 79.9% of the comments was found to be 15% with 18 cases in which it was achieved from a total of 120. By adding an 8th reviewer, the average coverage increases to 79.9% with an standard deviation of 8.36, a minimum coverage of 60%, maximum coverage of 95.6% and the probability of getting a set of reviewers that covers at least 80% becomes 26 ⁄ 45 = 0.5778 = 57.8%. Thus, we conclude a Minimum Set of 8 reviewers would be needed to cover the relevant comments in similar studies.

### Comment analysis: Classification and coding

This study followed a 4-cycle coding process to classify the comments provided by the reviewers in accordance to the suggested actions, displayed emotions and the overall topic treated in each comment. The process started with an in-vivo coding cycle to obtain the general categories of the comments, four categories were identified in this cycle:

* Requests for further explanation (“I would like this explained”)
* Requests for information (“I would like more information”, “They should be given details”, “(This) is not enough”)
* Request for clarity (“Too hard to understand”)
* Approval (“It was good”).

The next cycle classified the comments in accordance to their displayed emotions into positive, neutral and negative; an additional class (combination) was added for comments that contained emotions of different classes:

e.g. “Listing them is a good choice but not much is said about each procedure”.

The third cycle included motif coding, to group the comments based on the topic to which they referred. This coding technique was only applied to the comments on specific sections as the general comments proved to be too ambiguous to classify meaningfully e.g. “Not sure” and “It was good”. The motifs found in the third cycle were found to correspond with the topics found as relevant in assessing Patient Information Documents using the expanded EQIP scale (Charvet-Berard, Chopard, & Perneger, 2008) (The description of the study benefits, risks, procedure, purpose and possible impact on quality of life, the document design, the language employed in terms of tone and structure, the medical procedure and the treatment alternatives and finally the information sources and consent process).

The final cycle compared the data from the previous cycles and structured the overall classes as requests to change the sentence, requests to explain the terms, invitations to reflect on the information (cogitation), disagreements on the ideas expressed in the document (contention) and praises of the expressed ideas. The following sections describe in detail the findings of each cycle.

#### In-vivo coding

In the first instance the 240 comments were divided into general comments (m=85) and specific comments (n=149) based on whether they were associated with specific sections of the PILs or given as an overall description of the document quality. In the second step the comments were coded employing an in-vivo code to identify the general process described in the comments. As seen before, while most reviewers provided at least one general comment (n=23) only 42% provided any comments related to specific sections of the PILs.

The first class of comments were requests for further information. The participants commonly requested more information on specific elements of the document and thus most of the comments found in this class are associated with particular sections of the PIL. In the cases where a general comment was employed the comment included the specific element/information that should be expanded:

T1-OXI94: “I understand this word but not in detail so I would like this to be explained” – Associated section: “rehabilitation”

T2-Z217: “I feel that saying -insurance arrangements- is not enough in reassuring the people” –Associated section: “there are special insurance arrangements being put in place.”

T1-PAF84: “I think there is a need for amplifying some information during the procedure of providing general insight.” –Associated section: n/a

T1-Z215: “No info on what -rehabilitation- would involve generally” –Associated section: n/a

Another group of comments was composed of suggestions to change the content, structure or design of the document. As before, most of these comments were associated with specific sections of the documents:

T2-Z215: “<However> better word than <but>” –Associated section: “, but”

T2-FSS87: “Can this be explained better” –Associated section: “If you agree to take part in the study, your type of treatment will be chosen randomly by a computer”

T1-OXI94: “Can we not see it on a website? Or email for it?” –Associated section: “If you agree to take part in the study, your type of treatment will be chosen randomly by a computer”

T4-Z215: “FAQ not answered - will medication I take affect the study? Most older people are on several meds” –Associated section: n/a

Finally, the reviewers often employed general comments on give personal assessments of the quality of the document being reviewed. The comments in this group could be commending or highlighting general issues in the document design/information:

T3-GSN40: “Well, after I scanned all the PIL I did not find anything wrong about it. Well done” –Associated section: n/a

T2-Z215: “Need to number pages! I got completely lost” –Associated section: n/a

T3-SVG28: “Not sure” –Associated section: n/a

T1-KAY78: “It was excellent” –Associated section: n/a

T1-GTX38: “The whole patient information sheet did not give clear info; I am struggling to understand” –Associated section: n/a

A resume of the overall four classes of comments identified in this first cycle is given below:

Table In-vivo classes of comments

|  |  |  |
| --- | --- | --- |
| **In-vivo Code Class** | **Comment** | **Researcher Interpretation** |
| “I would like more information” | “I would like this explained”  “I understand this word but not in detail so I would like this to be explained”  “They should be given details”  “(This) is not enough” | This type of comment is a request for further explanation on specific terms, concepts, procedures or other information present in the document |
| “Instead of using ...” | “-However- better word than -but-”  “Simplify -relative merits- to advantages”  “Can we not see it on a website? Or email for it?” | These are indications to change the format, structure, or design of the reviewed document, procedure or information. |
| “Too hard to understand” | “Found it not so easy as the other two PILs asked to read. Found I had to go back and read several times. A lot of information to take in” | Mostly composed by general comments that give qualitative observations on the document information, composition and readability. |
| “It was good” | “I like the friendly writing style”  “This was a good one” | This are general comments that endorse the quality of the document. |

The classification process in this cycle determined that comments given by the reviewers were mostly concerned with asking for clarification on specific sections and requesting further information, or with assessing the quality of the documents and highlighting overall issues. These results were in accordance with the delineated tasks given to the reviewers. The general comments and the comments on specific sections were found to be grouped in different classes, with the general comments focusing on requesting further explanation and the general comments providing overall descriptions of the issues and qualitative assessments.

#### Emotion coding

In this cycle the emotions expressed by each comment were assessed into three categories, positive, neutral, negative. If a comment included emotions from more than one category, it was counted in each category it applied to. The tablesTable 5**,**Table 6summarize the emotive classes for comments given by the reviewers:

The positive category was mostly composed of general comments (49 general comments vs only 3 comments on specific sections). These comments included endorsements or commendations of the document, design or information:

T3-Z217: “All the details above were well organized, but the last bit informing of the harmless nature should have an emphasis on it” –Associated section: “Remember: Varicose veins are usually harmless and seldom cause serious medical problems”

T3-Z217: “Listing them is a good choice but not much is said about each procedure” –Associated section: “For people who want treatment there are three choices”

T4-RBM27: “Happy Face” –Associated section: n/a

T4-GSN40: “It seems very perfect and well organized. Good” –Associated section: n/a

The neutral category was composed mostly of comments associated with specific sections of the PILs (68 specific comments vs 6 general comments). These comments generally provided suggestions for improving the document quality which were not linked to the reviewer emotions:

T1-RBM27: “Grammar, it reads a bit awkwardly. Try -comparing day hospitals to rehabilitation at home- or vice-versa” –Associated section: “Rehabilitation for the elderly”

T1-SDF54: “what does randomised means?” –Associated section: “randomised”

T4-Z215: “May also help to say if preferred we can try to have the same researcher contact for Qs as may be less” –Associated section: n/a

T2-KXY38: “It is a better PIL than the first one” –Associated section: n/a

The negative section includes 88 specific and 26 general comments. These either provide a negative assessment of document quality, or highlight an emotive issue for the reviewer. 7 negative emotions were identified in this set of comments: Apprehension, Scepticism, Fear, Annoyance, Anger, Boredom and Confusion.

T1-NHB12-Confusion: “Too wordy and confusing” –Associated section: “so you will not be disadvantaged by being assigned to either one”

T2-KXY38-Boredom: “It was a bit long and boring”

T4-RBM27-Anger: “Long-term side effects factually not known of mentioned laxatives -be honest!” –Associated section: “However, like any medicines, laxatives can have unwanted side-effects in some patients such as abdominal “

T3-DBS58-Annoyance: “The PIL does not give all the info you need and it is not too specific” –Associated section: n/a

T1-FSS87-Apprehension: “Would a document need to be signed?” –Associated section: You have the option to withdraw at any time, for any reason”

T1-NHB12-Scepticism: “Why are they doing it then?” –Associated section: “From previous studies, we don’t expect there to be any difference in effectiveness between these two”

T1-OXI94-Fear: “I’d like to know who is doing the survey work/conversations. I need to think about keeping myself safe eg. Whether I want to invite a stranger into my home.” –Associated section: “I’d like to know who is doing the survey work”

Table Emotion classes for specific comments

|  |  |  |
| --- | --- | --- |
| **Emotion Code** | **Comment** | **Researcher Interpretation** |
| Positive | “Really good to inform where this is straight away as encourages person to come up with questions.”  “All the details above were well organized, but the last bit informing of the harmless nature should have an emphasis on it.” | Only 3 comments on specific sections were classified as positive from a set of 159. From those only 1 comment was deemed as purely positive. This was observed to be because these comments were used to give instruction about issues instead of qualitative assessments |
| Neutral | “This part should be underlined to be more clear”  “Inconsistent use of capitals”  “Seldom -> simplify for more people to understand easily -rarely-”  “Larger text, centralize or as a part of sentence (Still larger text)” | This type of specific comments was associated with instructions given to the researcher to further improve the document. They lacked any kind of emotive display. 68 comments from a set of 159 specific comments were deemed as neutral. |
| Negative | “Long-term side effects factually not known of mentioned laxatives -be honest!”  “Too subjective, best to whom or what end? -revise”  “Hmm! recommended for all at this stage? Seems risky” | This were the most numerous group of specific comments with 88 of 159 comments. These comments tended to confront the information present in the document. The emotions observed in this comments were: Apprehension, scepticism, fear, annoyance, anger, boredom and confusion. |

Table Emotion classes for general comments

|  |  |  |
| --- | --- | --- |
| **Emotion Code** | **Comment** | **Researcher Interpretation** |
| Positive | “This was a good one”  “I like the friendly writing style”  “Well, after I scanned all the PIL I did not find anything wrong about it. Well done.”  “The PIL is clear and well understood. It makes me to understand a lot of things I have never come across before.” | 49 general comments from a set of 81 were deemed as positive. These comments tended to express qualitative assessments of the documents. |
| Neutral | “May also help to say if preferred we can try to have the same researcher contact for Qs as may be less embarrassed to speak with one person rather than several”  “Advantages = getting treatment (injections) that the NHS may not afford” | Only 6 general comments were deemed to be neutral. They provided specific advice even when not associated with a particular section of the document. |
| Negative | “FAQ not answered - will medication I take affect the study? Most older people are on several meds”  “Add by contact info -if you have any other questions please use the contact info provided- to encourage active Qs/fb”  “The whole patient information sheet did not give clear info; I am struggling to understand.” | 26 general comments from a set of 81 were deemed as negative. These comments addressed issues within the document and were associated with annoyance and anger. |

The results from this coding cycle show that the emotions expressed in the comments vary greatly between comments associated with specific sections and general comments. This was observed to be a consequence of the use of both classes of comments. Specific comments were used to highlight particular issues in the document and thus were either neutral, when the reviewer just gave an instruction (“change seldom for rarely”), or negative, when the reviewer challenged a particular piece of information present in the document (“seems risky!”). On the other hand, general comments were commonly employed to give qualitative assessments on the overall quality of the document (“This was a good one”).

#### Motif coding

In this cycle, we employ motif coding to identify those inherent ideas that produced the comments and which recur in the different reviewed PILs. This cycle was applied only to the comments associated with specific sections of the PILs because the general comments tended to be too overreaching in their descriptions to properly code a motif. That is, the ideas behind the general comments were not found to be applicable to other documents, as most of them were simple personal quality assessments. The motifs found in this analysis were observed to correspond to the topics used by the Expanded EQIP scale (Charvet-Berard, Chopard, & Perneger, 2008) to determine the quality of patient information documents.

Our first motif class is composed of comments related directly to how the purpose of the study is explained to potential participants, whom reviewers highlighted that the way the statements were framed could be considered contradictory or induce distrust in the reader:

T1-NBH12: “why are they doing it then?” –Associated section: “From previous studies, we don’t expect there to be any difference in effectiveness between these two”

T1-FSS87: “If there have been previous studies, how come researchers still need to carry out new studies. Did they not gather this information sufficiently from these previous studies?” –Associated section: “From previous studies, we don’t expect there to be any difference in effectiveness between these two”

T2-FSS87: “Sounds unreassuring -if these surgery techniques really work-” –Associated section: “if these surgery techniques really work”

T1-ALT17: “If there have been previous studies, how come researchers still need to carry out new studies. Did they not gather this information sufficiently from these previous studies?” –Associated section: “From previous studies, we don’t expect there to be any difference in effectiveness between these two”

Our next motif class stems from the description of the benefits, risks and disadvantages provided by the study to the potential participants, which are considered lacking or badly explained. The reviewers gave higher emotive responses to the comments in this group:

T3-Z217: “Give people a reason to join the study” –Associated section: “Introduction to the study”

T2-Z217: “A more clear approach to the advantages as to encourage people to consider the study” –Associated section: “Advantages”

T1-FSS87: “People will need to know if there are risks or disadvantages not -we don’t think there are- this is an open statement” –Associated section: “We don’t think there are any risks or disadvantages”

T1-OXI94: “I’d like to know who is doing the survey work/conversations. I need to think about keeping myself safe e.g. Whether I want to invite a stranger into my home” –Associated section: “We don’t think there are any risks or disadvantages for being involved in this study”

Another major motif group was about how the study procedures are explained. The comments in this group criticised the lack of detail when referring to how the study was to be approached. This also applied to descriptions and justifications of the elements of medical procedures when they were required by the study:

T1-OXI94: “How will I be updated? Do I have to make the effort to contact you” –Associated section: “There will be an opportunity for you to see the results of the study when it is completed”

T2-Z217: “Explain why a person would be chosen to join the trial” –Associated section: “Why have I been approached?”

T2-FSS87: “Is general anaesthetic necessary?” –Associated section: “This procedure requires a general anaesthetic.”

Comments which referred to the language employed in the document formed the largest motif group. This included comments on the use of jargon, repetitive language, bad grammar and punctuation, ambiguous phrasing and the tone of the document:

T2-OXI94: “A big word to put at the beginning” –Associated section: “Arthroscopic lavage”

T2-Z215: “No comma needed” –Associated section: “creams, steroid injections, and surgery”

T1-RBM27: “Could be a bit abrupt? Should not the patient decide their needs? Rewards?” –Associated section: “You have been chosen because you are an elderly person who has been identified as needing rehabilitation”

T1-NHB12: “Wordy” –Associated section: “The study is a National Randomised Controlled Trial, which means it is taking place nationally”

T4-RBM27: “Repeating in different words too much -find simpler way of explaining the fact” –Associated section: “We want to reassure you that anything you tell us will be kept secret. We will not tell anyone what you have said unless you ask us to. We will not give your contact details to anybody and nobody else will contact you by any means after the end of the study.”

T1-FSS87: “What was the result of previous studies & how long ago?” –Associated section: “From previous studies, we don’t expect there to be any difference in effectiveness between these two”

T1-NHB12: “Rehab for what?” –Associated section: “you are an elderly person who has been identified as needing rehabilitation”

Other motif topics which included the comments of only one or two reviewers were: document design, study impact on the patient quality of life, the description of the treatment alternatives, the presentation of insurance details, the inclusion of sources of information and the consent process:

T1-Z217: “Put a list of any differences between the methods” –Associated section: “The study aims to find any differences”

T2-Z217: “This part should be underlined to be more clear” –Associated section: “Please take time to decide whether or not you wish to take part”

T1-OXI94: “Not clear, because the rehabilitation will make changes to my lifestyle but the talking to someone about it will not” –Associated section: “Taking part in the study does not require you to make any changes to your lifestyle”

T3-Z217: “That is quite vague, in that situation I would like to be more clear on the effect of the study on my personal life” –Associated section: “It requires approximately 2 to 3 days off work”

T3-RBM27: “Misleading there are no doubt other treatments available not on the NHS” –Associated section: “three choices”

T2-Z217: “I feel that saying -insurance arrangements- is not enough in reassuring the people” –Associated section: “there are special insurance arrangements being put in place”

T1-OXI94: “I would like this to be on a website or email as it is a lot to ask people to write a letter” –Associated section: “a copy may be obtained from CERES”

T3-Z217: “The paragraph about more information should be clearly marked. I see it hidden as a side note, but in truth it is important” –Associated section: “If you want more information “

T1-NBH12: “How?” –Associated section: “If you decide to take part you will be given this information sheet to keep and be asked to sign a consent form”

T1-FSS87: “Would a document need to be signed?” –Associated section: “You have the option to withdraw at any time, for any reason”

#### Final coding

Based on the results of the previous coding cycles, a final coding framework was developed that accounts for both the function of the comments and the displayed emotion to create five classes. These classes were: requests to change a sentence, requests to explain a term, the induction of reflection, the endorsement of the information and contention about ideas expressed in the document.

The first class of comments are statements that endorse the ideas, structure and presentation of the documents. These comments are generally positive assertions related to the overall quality of the reviewed documents:

T3-Z217: “All the details above were well organized, but the last bit informing of the harmless nature should have an emphasis on it” –Associated section: “Remember: Varicose veins are usually harmless and seldom cause serious medical problems”

T3-Z217: “Listing them is a good choice but not much is said about each procedure” –Associated section: “For people who want treatment there are three choices”

T4-RBM27: “Happy Face” –Associated section: n/a

T4-GSN40: “It seems very perfect and well organized. Good” –Associated section: n/a

The second group of comments is composed by emotively neutral brief requests to change specific terms, words or phrases. They could directly propose an alternative wording or just give a general indication of the need to restructure the section:

T2-Z215: “No comma needed” –Associated section: “creams, steroid injections, and surgery”

T1-NHB12: “Too wordy and confusing” –Associated section: “so you will not be disadvantaged by being assigned to either one”

T2-Z215: “<However> better word than <but>” –Associated section: “, but”

The next group of comments are requests to explain specific terms in more detail. This comments are emotively neutral:

T4-Z215: “A quick overview of what the health diary involves may be useful” –Associated section: “Health Diary”

T3-OXI94: “I don’t know what this means” –Associated section: “duplex”

T1-SDF54: “explain -future policy development-” –Associated section: “future policy development”

The following group of comments invite the researcher to reflect on the information presented in the PIL (cogitation). They can display emotions like confusion and apprehension but are not statements that refute, contradict or question the validity of the claims made by the information:

T2-SDF54: “who are the appropriate agencies?” –Associated section: “appropriate agencies”

T1-RBM27: “Could be a bit abrupt? Should not the patient decide their needs? Rewards?” –Associated section: “You have been chosen because you are an elderly person who has been identified as needing rehabilitation”

T1-OXI94: “How will I be updated? Do I have to make the effort to contact you” –Associated section: “There will be an opportunity for you to see the results of the study when it is completed”

T2-FSS87: “Is general anaesthetic necessary?” –Associated section: “This procedure requires a general anaesthetic.”

T1-NHB12-Confusion: “Too wordy and confusing” –Associated section: “so you will not be disadvantaged by being assigned to either one”

T1-OXI94: “I’d like to know who is doing the survey work/conversations. I need to think about keeping myself safe e.g. Whether I want to invite a stranger into my home.” –Associated section: “I’d like to know who is doing the survey work”

The final group of comments are highly emotive and can be based on specific PIL sections or be general assessments of the text quality. They call into question the validity, relevance and completeness of the presented information and in some cases contradict them:

T1-FSS87: “People will need to know if there are risks or disadvantages not -we don’t think there are- this is an open statement” –Associated section: “We don’t think there are any risks or disadvantages”

T3-KXY38: “It was a bit long and boring”

T4-RBM27: “Long-term side effects factually not known of mentioned laxatives -be honest!” –Associated section: “However, like any medicines, laxatives can have unwanted side-effects in some patients such as abdominal “

T3-DBS58: “The PIL does not give all the info you need and it is not too specific” –Associated section: n/a

T1-NHB12: “Why are they doing it then?” –Associated section: “From previous studies, we don’t expect there to be any difference in effectiveness between these two”

The comments analysed in this section were found to have a clear correspondence with the identified topics in the Expanded EQIP scale (Charvet-Berard, Chopard, & Perneger, 2008) as important when assessing the quality of a PIL. Some of the topics found in the comments but not assessed with the Expanded EQIP scale include the consent and randomization processes, the inclusion of irrelevant or repeated information and the use of bad writing (grammar, spelling, punctuation or inappropriate language). As the reviewers were presented only with the anonymised PIL text some topics present in EQIP were not commented on, including the presence of the date of issue, logos, names of persons/entities who created and financed the study and the inclusion of a consent form.

### Association between the percentage of correct answers, PIL quality grade and the comments provided by the reviewers.

A simple linear regression model was employed to assess the association between the subjective PIL quality grade given by the reviewer and the percentage of correct EQIP answers they obtained. The model found a significant association between the variables with *R*2 = 0.086 and equation *F*(1, 71) = 6.87,  *p* < 0.012,  *y* = 28.9 + 3\**QualGrade*.

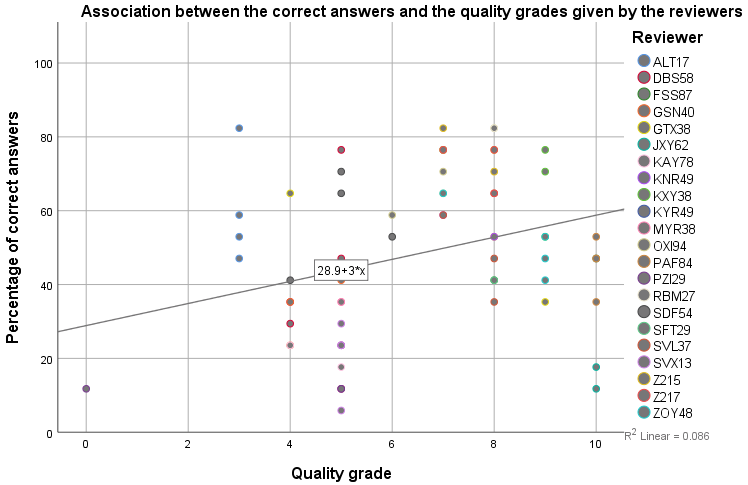


Figure Association between PIL quality grade and the percentage of correct EQIP answers

The model discovered a significant association between the percentage of correct answers obtained by the reviewers and the quality grades they provided. However, the percentage of correct answers only accounts for 8.6% of the variance in the quality grades.

Univariate regression models were employed to assess the association between the percentage of correct answers and of the number of general comments and comments given to specific section of the PILs. No significant association was found between the number of general comments and the percentage of correct answers (*p* = 0.247) or the reviewers’ quality grades (0.229), but a significant association between the number of comments given to specific section of the PILs was identified.

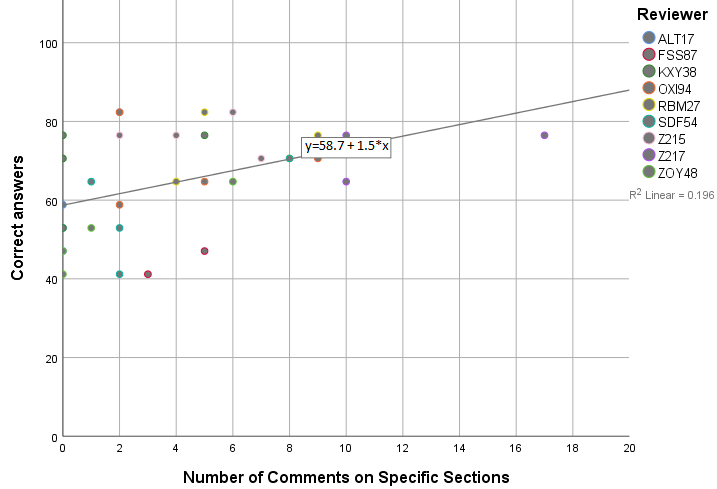


Figure Association between the number of comments on specific sections and the percentage of correct EQIP answers obtained by the reviewer

The linear regression formula to represent the association between the number of specific comments and the percentage of correct answers was defined as *F*(1, 31) = 7.6,  *p* < 0.01,  *R*2 = 0.196 and *y* = 58.7 + 1.5*SpecComments*. Finally, no significant association was found between the number of specific comments and the reviewers’ PIL quality grades (*p* = 0.456).

### Demographic factors associated with the PIL quality grades

The level of correlation between the ordinal demographic variables and the quality grades given by the reviewers to the PILs was measured with a Pearson correlation model (Figure 17). The variables that presented moderate or strong correlations and p-values less than 0.02 were included in a multivariate regression model**.** Age group (*rp* =  − 0.665,  *p* < 0.001) and health level (*rp* =  − 0.396 *p* < 0.001) were identified as having meet the criteria.

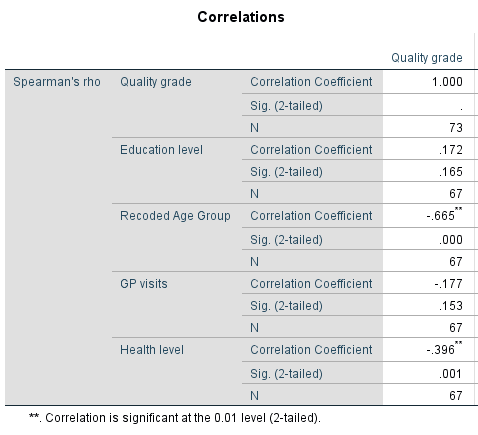


Figure Ordinal demographic factors correlated with the PIL quality grade

A visual inspection of the nominal variable boxplots (Figure 18,Figure 19,Figure 20) was used to identify nominal variables with a possible association with the quality grades. Dummy variables were created for the identified factors and added to the multivariate regression model. The identified factors were: interest in participating in RCTs, gender and origin.

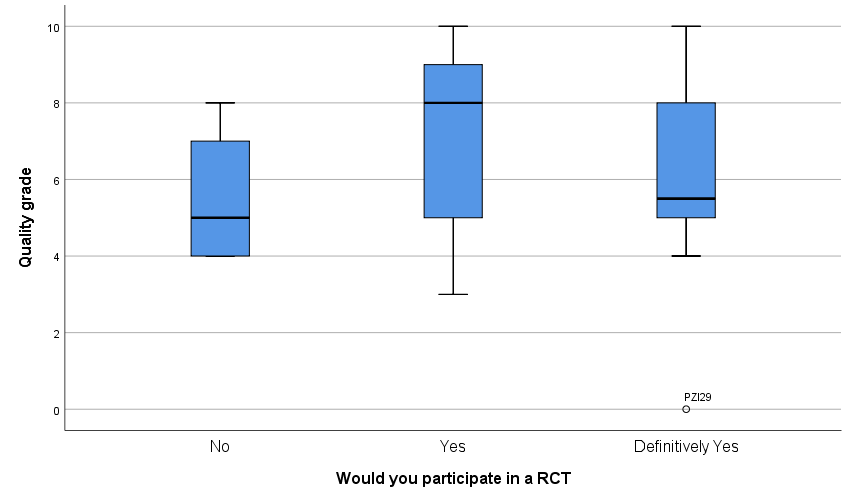


Figure [Boxplot] Quality grades grouped by interest in participating in an RCT

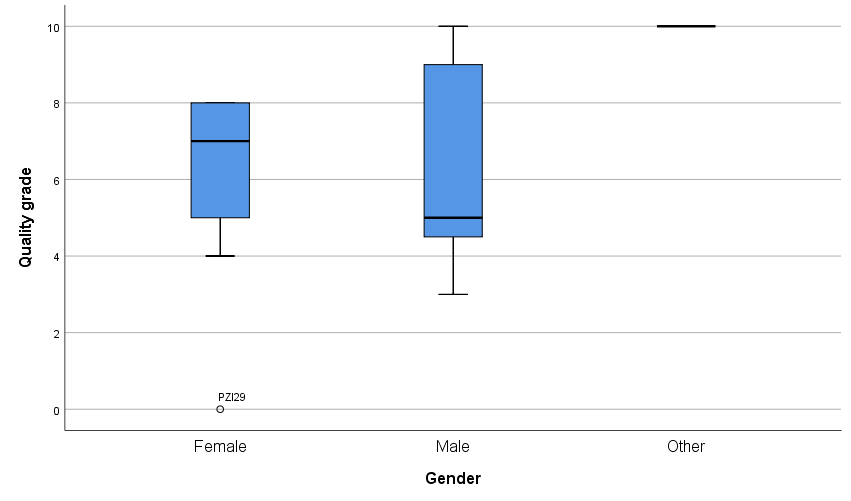


Figure [Boxplot] Quality grades grouped by gender

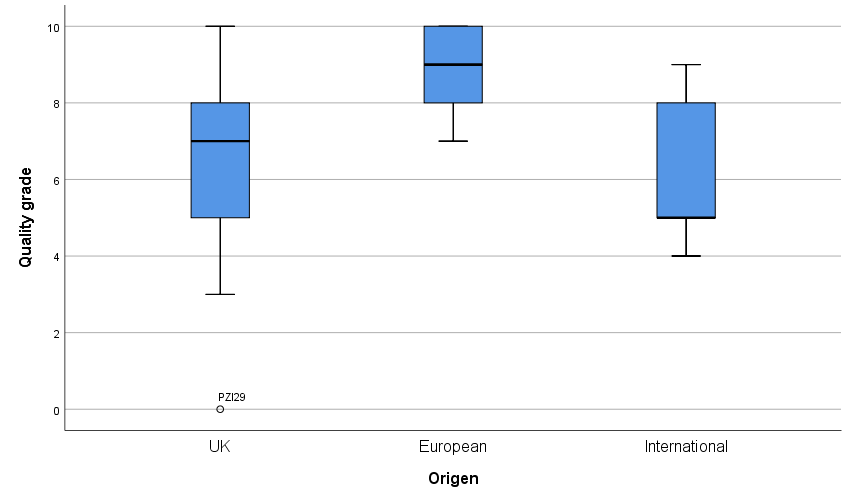


Figure [Boxplot] Quality grades grouped by origin

Table Factors included in the multivariate linear regression model

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Variable** | **Type** |
| Origin: UK | Nominal | Interest: No | Nominal |
| Origin: European | Nominal | Interest: Yes | Nominal |
| Origin: International | Nominal | Interest: Definitively | Nominal |
| Gender: Male | Nominal | Age group | Ordinal |
| Gender: Female | Nominal | Health level | Ordinal |

The level of association of the previously identified factors Table 7 was assessed by employing a multivariate linear regression model with the stepwise method in SPSS with the formula *F*(3, 60) = 33.4,*p* < 0.001 and *R*2 = 0.626. **Three** final factors with statistically significant associations were identified in this data: **age group** (*B* =  − 2.2), **health level** (*B* =  − 1.1) and **reviewers who indicated they would not participate in an RCT** (*B* =  − 1.4). The linear equation was defined as:

*y* = 14.8 − 2.2*AgeGrp* − 1.1*HealthLvl* − 1.4*PartNo*

### Demographic factors associated with the reviewers’ understanding of trial features

A bipartite Pearson correlation model was used to determine which ordinal demographic variables were associated with the percentage of correct EQIP answers obtained by the reviewerFigure 21. The variables with moderate and strong correlations (*rs* ≥ 0.3 or *rs* ≤  − 0.3) and p-values less than 0.02 were education level, age group and health level. These were added to the multivariate regression model.

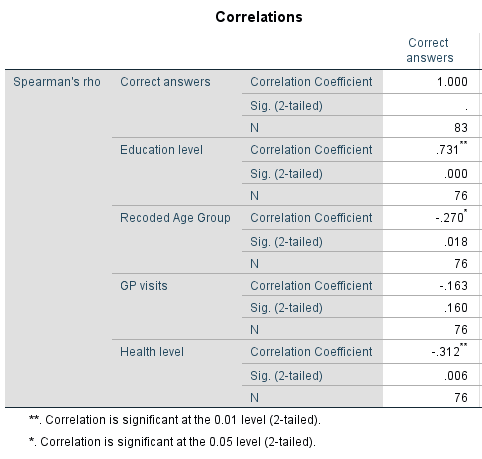


Figure Ordinal demographic variables associated with the percentage of correct EQIP answers

A visual inspection of the boxplots for the nominal variables identified gender, origin and interest in participating in an RCT as possible factors associated with the percentage of correct answersFigure 22**,**Figure 23**,**Figure 24. The respective dummy variables were generated for these factors to be included in the multivariate linear regression model.

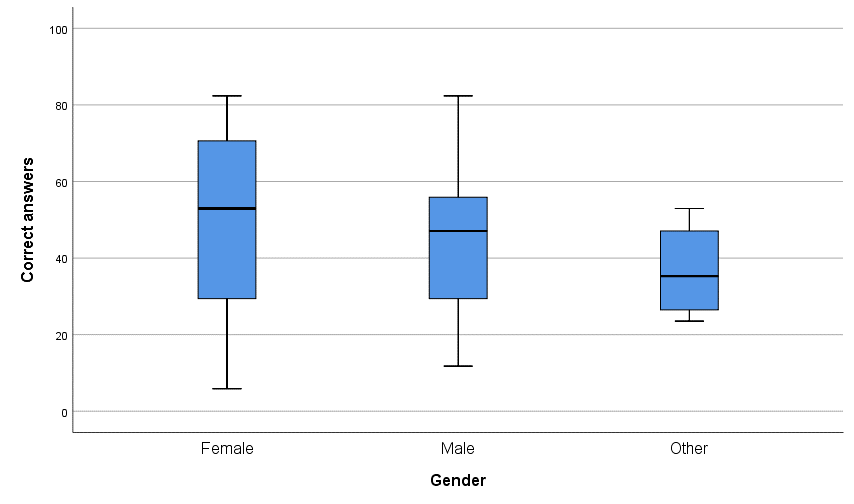


Figure [Boxplot] Percentage of correct EQIP answers grouped by gender

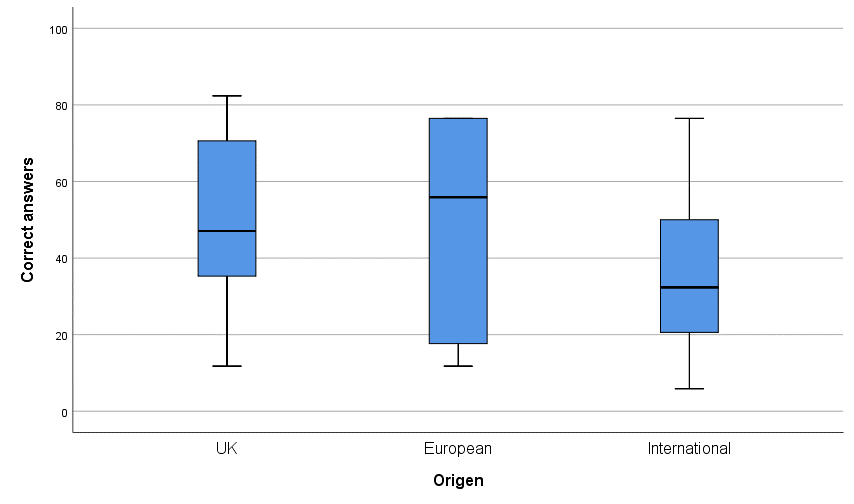
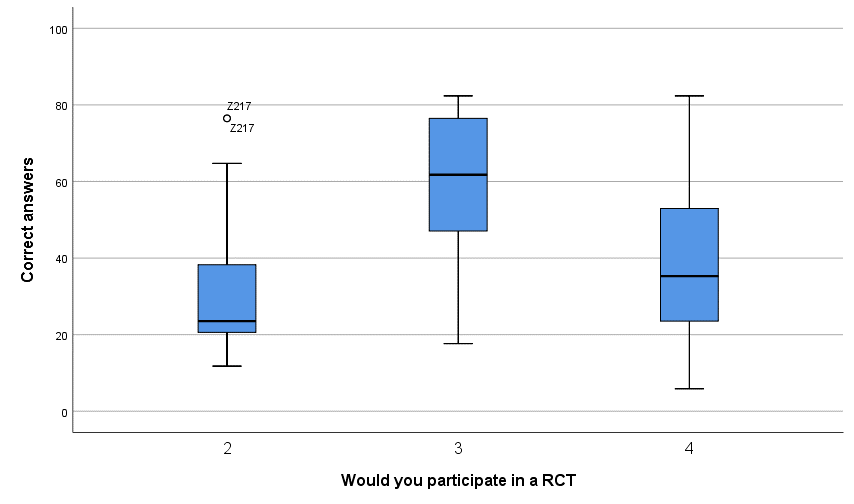


Figure [Boxplot] Percentage of correct EQIP answers grouped by origin



Definitively

Yes

Figure [Boxplot] Percentage of correct EQIP answers grouped by interest in participating in an RCT

No

The previous identified possible demographic factorsTable 8associated with the percentage of correct answers obtained by the reviewers were added to a multivariate linear regression model with the stepwise method in SPSS.

Table Input variables for the linear regression model

|  |  |
| --- | --- |
| **Variable** | **Type** |
| Gender | Nominal |
| Origin | Nominal |
| Preference to participate in an RCT | Nominal |
| Education level | Ordinal |
| Age group | Ordinal |
| Health level | Ordinal |

Those factors were added into a single multivariate linear regression model with the formula *F*(3, 68) = 32.2,*p* < 0.001 and *R*2 = 0.59. Three final factors with statistically significant associations to the percentage of correct answers were identified in this study, education level (*B* = 8.8), age group (*B* =  − 5.6) and health level (*B* =  − 8.0). The linear equation was defined as:

*y* = 53.8 + 8.8*EdLvl* − 5.6*AgeGrp* − 8*HealthLvl*

## Validation analysis: Lexicon polarity vs manual annotation

A comparison of the emotion codes manually annotated in the previous analysis was made with the automated labels obtained from running the NRC Emotion Lexicon (EmoLex) (Mohammad & Turney, NRC emotion lexicon, 2013) and the NLTK sentiment polarity algorithm (Pang & Lee, 2004) over the comment text. 240 manually annotated comments were compared with the outputs of the previous algorithms. Seven comments were excluded from this analysis as they were deemed to express a mixture of positive and negative sentiment or did not contain text (e.g. a happy face):

T3-z217: “All the details above were well organized, but the last bit informing of the harmless nature should have more emphasis” –Associated Section: “Remember: Varicose veins are usually harmless and seldom cause serious medical problems”.

Two of the excluded comments were associated with specific sections of the PILs while the other five were general comments, these represented 1.3% and 5.9% of the total comments.

Table Distribution of the comments annotated by the NLTK

|  |  |  |  |
| --- | --- | --- | --- |
|  | Specific comments | General comments | Total Comments |
| NLTK positive comments | 43 | 25 | 68 |
| NLTK neutral comments | 66 | 7 | 73 |
| NLTK negative comments | 48 | 44 | 92 |
| Manual annotated positive comments | 1 | 44 | 45 |
| Manual annotated neutral comments | 68 | 6 | 74 |
| Manual annotated negative comments | 88 | 26 | 114 |

Table Distribution of the labels for specific comments

|  |  |  |  |
| --- | --- | --- | --- |
|  | Manual (Specific) + | Manual (Specific) 0 | Manual (Specific) - |
| NLTK + | 34 | 23 | 0 |
| NLTK 0 | 4 | 32 | 8 |
| NLTK - | 6 | 17 | 31 |

Table Distribution of labels for general comments

|  |  |  |  |
| --- | --- | --- | --- |
|  | Manual (General) + | Manual (General) 0 | Manual (General) - |
| NLTK + | 3 | 0 | 29 |
| NLTK 0 | 0 | 0 | 29 |
| NLTK - | 0 | 0 | 19 |

The distribution of the labels the NLTK algorithm assigned was found to differ greatly from the annotated labels, with only 41% of the labels coinciding for comments associated with specific sections and 66% accuracy on the general comments. Another analysis was done employing the NRC Emotion Lexicon (Mohammad & Turney, NRC emotion lexicon, 2013), this lexicon provides the associations to principal emotion categories (Plutchik, 1984).

Table Distribution of comments annotated using the NRC Emotion Lexicon (EmoLex)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Specific comments | General comments | Total Comments |
| [EmoLex] Comments with positive words | 47 | 38 | 85 |
| [Emolex] Comments without positive or negative words | 92 | 29 | 121 |
| [EmoLex] Comments with negative words | 11 | 4 | 15 |
| [EmoLex] Comments with positive and negative words | 9 | 10 | 19 |
| Manual annotated positive comments | 1 | 44 | 45 |
| Manual annotated neutral comments | 68 | 6 | 74 |
| Manual annotated negative comments | 88 | 26 | 114 |
| Manual annotated positive and negative comments | 2 | 5 | 7 |

The lexicon found words related to positive sentiment in 47% of the manually annotated positive comments, words related to negative sentiment in 9.6% of the manually annotated negative comments and determined no emotive words were present for 64% of the comments manually annotated as neutral.

## Conclusions

While there exists a significant association between the subjective quality grades and the understanding of the trial features (percentage of correct EQIP answers), the quality grades only account for 8% of the variance in the percentage of correct answers obtained by the reviewers. Even more, when the number of specific comments per PIL given by the reviewers is added to the linear regression model (*R*2 = 0.22,  *p* < 0.007) the quality grades are deemed non-significant (*p* < 0.682). This illustrates the need to independently assess the level of understanding of the trial features by members of the public, as it cannot be reliable inferred from other data.

Our research found significant negative associations between the age group of the reviewers, the reviewer self-assessed health level and the perceived quality of the PILs (measured by a unipolar scale 0-10). Also, those reviewers who indicated a lack of interest in participating in RCTs had a significant tendency to give harsher scores to the PILs. Our final linear regression model accounted for 62% of the variance in the quality grades. This shows the need to consider carefully not only the age of the reviewers but also rarely recorded factors like the participant’s interest in RCTs and how healthy they perceive themselves to be.

The analysis of factors significantly associated with the reviewers understanding of the trial features (percentage of correct answers to EQIP quiz) showed that participants with higher levels of education tend to score better. It also shows that older participants and participants who perceive themselves as unhealthy tend to have more difficulty understanding the information in the PILs. These results account for 58.7% of the variance in the percentage of correct answers given by the participants, which indicates a concerning tendency for low education, elder and unhealthy patients to have trouble understanding the information presented to them, who often are the people targeted for trial recruitment.

Most of the comments which were significantly associated with either the reviewer understanding of the trial features or their perception of the PIL quality were about topics which correlated with the Expanded EQIP scale. The topics mentioned by reviewers which were not found in the scale refer to the processes of consent and randomisation and the use of irrelevant, badly written (grammar, spelling, punctuation or inappropriate terms) or repeated information.

Finally, the comparison between automated labelling by the NLTK and NRC EmoLex lexicons with the manual annotation of the sentiment polarity of the comments, showed substantial divergence in all the cases. This has taken us to consider necessary further verification of the validity of the lexicons. Thus, a sentiment-emotion labelling function of the reviewer comments has been added to the Web platform for the next study.

# Assessing the feasibility of using a Web Platform and Amazon Crowdsourcing to Revise the Text of RCT PILs.

## Research Questions

1. Is it feasible to use a Web platform to collect, link, store and present public feedback on the text of RCT PILs?
2. Is it feasible to employ a Web platform to employ Amazon Turk workers to revise the wording of PIL sentences that are too difficult to understand?
3. Is there evidence that giving revisers tips on how and what to revise enables them to offer revised sentences that are easier to understand?
4. Which is the optimal number of sentences that each reviser should be given to ensure the revisions quality?

## Aims

1. To measure the PIL authors response to employing a Web platform for collecting, associating, analysing, and presenting the public feedback on their PILs.
2. To assess the learning and fatigue effects on Amazon participants when rewriting PIL sentences that are too hard to understand.
3. Identifying the demographics and minimum health literacy skill required for the participants to be able to revise PIL sentences.

## Background & Rationale

Our previous research has found that PILs are quite different from the texts average people read day to day (Master Thesis). We also found that current PILs do not provide the reader with information that is easy to understand. Furthermore, based on discussions with Patient and Public Involvement officers, Principal Investigators (PI) generally do not evaluate how well the patient comprehends the essential aspects of the trial before signing consent. Several independent studies have implicated that the current PILs are badly written and cannot meaningfully inform the patient.

This is not to say that no interest is given to the design of PILs: UK regulation establishes their purpose as “should support the consent process by helping to ensure that all those who are invited to take part in a research study have been adequately informed” (HRA, 2017), and ensures that every trial which recruits patients should create patient information that is understandable by the patients and take their circumstances into account. Furthermore, several of UK research funders would ask the principal investigator to collaborate with a Patient and Public Involvement (PPI) group in a review process to ensure information quality before granting the research funds.

Recent research like the Health Research Board Trials Methodology and Networks (TMRN) collaboration with the James Lind Alliance and the TrialBank have identified several high-priority questions relating to the effects public reviewers have on PILs and RCTs (Healy, et al., 2018). In particular, focus has been given to identifying which information should be given, finding the best designs to deliver the information and assessing the effect of PPI collaboration on recruitment (Healy, et al., 2018).

With this in mind, this research seeks to evaluate several techniques employed in analysing Web text and linking information based on relationship models. We propose to evaluate the use of sentiment and content analysis in conjunction with a relationship database model and clustering to provide a Web platform in which PI and public reviewers could interact as a form to measure the quality of the text, identify issues and easily visualise the effects of PPI.

## Methods

This study is composed from 5 separate experiments:

### Employing a Web platform to collect public feedback on PIL text.

A personalized website was created were public participants were presented the text of 4 PILs

This study is a randomised experiment with 5 non-medical interventions with independent group. The groups are composed of Principal Investigators (PIs) who are presented with feedback from public reviewers and the selected Web technique on their previous PILs.

#### Sample

24 public revisers

#### Recruitment

Flyers were posted in public locations inviting participants to attend a meeting at The Millennium Third Age Centre (3AC). The 3AC is a charity providing community resources and delivering community activities to Southampton. Participants were asked to read and comment the text of 3 PILs in a Website, and were offered up to £10 for their work.

#### Procedure

The participants would enter a website to register and join the study after reading the study details. In a first task they were asked to read and comment on the text of 3 PILs, the comments were made by selecting sections of the PIL text presented in the Web page clicking a button to make a comment and filling a comment form. A second task asked the participants to fill the extended EQIP questionnaire to assess the quality of the PIL information. Finally, each participant was asked to rate the quality of the information employing a unipolar scale from 0-10 and to determine if the PIL was usable, required minor corrections before use or needed to be redrafted.

#### Analysis and Measures

The analysis of this experiment was focussed on the type of comments received for the PILs and the correlation between the comments, the PIL quality perception of the participants, the EQIP quantitative score for the PIL information quality and the quantitative readability metrics of the PIL text.

### Using the Cloze procedure to identify jargon terms and structural problems in the PIL sentences.

A complete feasibility study on employing the Cloze procedure to identify jargon words and syntactical issues on PIL sentences

#### Sample

45 Adult Amazon Turk revisers who reside in the UK.

#### Recruitment

A Human Intelligence Task will be posted in the Amazon Mechanical Turk platform inviting the public to participate in revising 3 PIL sentences. A compensation of $1 usd will be offered to those who complete the task

#### Procedure

A HIT was posted for each of the experiment arms, each HIT invited 9 participants to revise 3 PIL sentences. Before revising the sentences, the participants were asked to fill complete the sentences via the Cloze procedure to assess both the difficulty of the sentence and the health literacy skill of the participant. One fifth of the words in the sentences were replaced by the Cloze procedure in each of the arms. These words were selected with a staggered start design, i.e. the first arm would replace the 1st, 5th… words while the second arm replaced the 2nd, 6th and so on.

#### Analysis and Measures

The participant performance in terms of sentence difficulty, time spent on each tasks, the percentage of correctly identified words and the readability score of the proposed revisions will be assessed to identify if the Cloze procedure produces consistent observations that permit to identify jargon words or syntactical issues on the PIL sentences.

### Identifying the required health literacy skill needed to revise PIL sentences

A complete feasibility study on employing the Cloze procedure to assess the readability of PIL sentences and the health literacy skill of Amazon revisers.

#### Sample

54 Adult Amazon Turk revisers who reside in the UK.

#### Recruitment

A Human Intelligence Task will be posted in the Amazon Mechanical Turk platform inviting the public to participate in revising 3 PIL sentences. A compensation of $1 usd will be offered to those who complete the task

#### Procedure

A HIT was posted for each of the experiment arms, each HIT invited 9 participants to revise 3 PIL sentences. Before revising the sentences, the participants were asked to fill complete the sentences via the Cloze procedure to assess both the difficulty of the sentence and the health literacy skill of the participant. Each participant will review the same sentences and with the same omitted words in by the Cloze procedure. Each of the arms will present the sentences in a different order based on the difficulty of a particular sentence (Easy, Medium or Hard).

#### Analysis and Measures

The difficulty of a sentence is determined by the average grade of four readability indexes, ARI, Gunning-Fog, Flesch-Kincaid and SMOG. A sentence is considered Hard when it requires a professor level reading skill, Medium at graduate level, and Easy bellow graduate level. All selected sentences have readability scores above 9th grade which is the recommended level for public audiences.

The participant performance in terms of sentence difficulty, time spent on each tasks, the percentage of correctly identified words and the readability score of the proposed revisions will be assessed to identify if the Cloze procedure produces consistent observations that permit to assess the readability of PIL sentences.

### Assessing the learning and fatigue effects when revising PILs sentences

We plan to conduct a complete randomised feasibility study of Amazon Turk as a platform for crowdsourcing the revision of PIL sentences that are too hard to understand.

#### Sample

We plan to recruit 108 Adult Amazon Turk workers who reside in the UK.

#### Recruitment

The participants will be recruited by posting a Human Intelligence Task (HIT) on the UK Amazon Turk platform. A reward of $1 usd will be granted for the completion of the task.

#### Procedure

The study will be composed of 9 arms, each of the arms will present the same tasks to the revisers but in different order.

The tasks consist of rewording 9 sentences that have been determined require a reading skill above 9th grade.

The order of the tasks will be fixed within each branch. The study will have a staggered start, i.e. Participants in the 1st branch will start by rewriting the first sentence, while participants in the 2nd branch will start with the second sentence and so on.

#### Analysis and Measures

The necessary reading skill to understand each sentence has been assessed by employing an agreement algorithm between the scores of four commonly used readability indexes ARI, SMOG, Gunning-Fog, and Flesch-Kincaid.

The algorithm evaluated the Coefficient of Variation between the reported scores for each of the readability indexes. Sentences with coefficient of variation above 0.2 were excluded. A selection of 9 sentences with average readability scores between 10th grade and 12th grade was made.

The randomization of the participants will be done by an automated PHP algorithm in the website. The algorithm will randomly assign the participant to one of the branches which have not yet completed their recruitment at the start.

The performance of the revisers will be assessed by observing the time taken to complete each task and the readability score of the proposed revised sentence. The readability of the revised sentences will be obtained by employing the procedure previously described.

### Assessing the viability of the proposed revisions under the PIL authors perspectives

A complete feasibility study on employing the Cloze procedure to identify jargon words and syntactical issues on PIL sentences

#### Sample

3 PIL authors.

#### Recruitment

PIL authors for clinical trials associated to the University of Southampton have been invited to participate in Public Involvement workshops to give input on the Web platform design and functions. Authors for the Macmillan HORIZONS Study, the Evaluation of the impact of a Prostate Cancer Survivorship Care Programme, and the TrueNTH Global Registry Participant study have accepted to be involved.

#### Procedure

A set of 27 revised sentences for their PILs will be offered to each author. The set will be composed of three revisions for 9 original sentences in the PILs. The authors would be asked to assess the quality of each of the proposed revisions in terms of their coverage of the original content, the ease for communicating the information, the relevance of the sentence to the trial information, and the appropriateness of the tone for a medical document.

#### Analysis and Measures

The participants’ assessments will be compared to the quantitative readability scores for each sentences, and the ranking given to each sentence on their appropriateness for use by a public group.

# Research Schedule



# Bibliography

Adams, H. F. (1916). *Advertising and its mental laws.* The Macmillan Company.

Averill, J. R. (1980). On the paucity of positive emotions. In *Assessment and modification of emotional behavior* (pp. 7-45). Springer.

Blaney, P. H. (1986). Affect and memory: a review. *Psychological bulletin, 99*, 229.

Bollobás, B. (1998). Modern graph theory, volume 184 of Graduate Texts in Mathematics. Springer-Verlag, New York.

Bolls, P. D., Muehling, D. D., & Yoon, K. (2003). The effects of television commercial pacing on viewers' attention and memory. *Journal of Marketing Communications, 9*, 17-28.

Bondy, J. A., & Murty, U. S. (1976). *Graph theory with applications* (Vol. 290). Citeseer.

Bower, G. H. (1992). How might emotions affect learning. *The handbook of emotion and memory: Research and theory, 3*, 31.

Bower, G. H., & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science, 247*, 44-49.

Charmaz, K. (2014). *Constructing grounded theory.* Sage.

Charvet-Berard, A. I., Chopard, P., & Perneger, T. V. (2008). Measuring quality of patient information documents with an expanded EQIP scale. *Patient education and counseling, 70*, 407-411.

Clements, F. E. (1954). Use of cluster analysis with anthropological data. *American Anthropologist, 56*, 180-199.

Darwin, C. (1872). The expression of the emotions in man and animals, New York: D. *Appleton and Company*.

Darwin, C. (1998). *The expression of the emotions in man and animals.* Oxford University Press, USA.

DeWall, C. N., & Myers, D. G. (n.d.). Words Have the Power to Inspire Violence and Peace.

Dixon, T. L., & Linz, D. (2002). Television news, prejudicial pretrial publicity, and the depiction of race. *Journal of Broadcasting \& Electronic Media, 46*, 112-136.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational research methods, 10*, 5-34.

Eighmey, J., & Sar, S. (2007). Harlow Gale and the origins of the psychology of advertising. *Journal of Advertising, 36*, 147-158.

Ekman, P. (1973). Cross-cultural studies of facial expression. *Darwin and facial expression: A century of research in review*, 169-222.

Ekman, P. E., & Davidson, R. J. (1994). *The nature of emotion: Fundamental questions.* Oxford University Press.

Entman, R. M., & Rojecki, A. (2001). *The black image in the white mind: Media and race in America.* University of Chicago Press.

Escudero-Carretero, M. J., S{\'a}nchez-G{\'o}mez, S., Gonz{\'a}lez-P{\'e}rez, R., Sanz-Amores, R., Prieto-Rodr{\'\i}guez, M. A., & Fern{\'a}ndez de la Mota, E. (2013). Elaboración y validación de un documento informativo sobre adeno-amigdalectomı́a para pacientes. *Anales del sistema sanitario de Navarra*, *36*, pp. 21-33.

Eysenck, M. W. (2012). *Simply psychology.* Psychology Press.

Gillies, K., Huang, W., Skea, Z., Brehaut, J., & Cotton, S. (2014). Patient information leaflets (PILs) for UK randomised controlled trials: a feasibility study exploring whether they contain information to support decision making about trial participation. *Trials, 15*, 62.

Hart, G. (2002). The five W’s of online help systems.

Healy, P., Galvin, S., Williamson, P. R., Treweek, S., Whiting, C., Maeso, B., . . . others. (2018). Identifying trial recruitment uncertainties using a James Lind Alliance Priority Setting Partnership--the PRioRiTy (Prioritising Recruitment in Randomised Trials) study. *Trials, 19*, 147.

Henry, D., Dymnicki, A. B., Mohatt, N., Allen, J., & Kelly, J. G. (2015). Clustering methods with qualitative data: a mixed-methods approach for prevention research with small samples. *Prevention Science, 16*, 1007-1016.

Hilton, M. (2007). Consumers and the state since the Second World War. *The Annals of the American Academy of Political and Social Science, 611*, 66-81.

HRA. (2017, January 17). Appliying a proportionate approach to the process of seeking consent - HRA Guidance.

Jansson-Boyd, C. V. (2010). *Consumer psychology.* Open University Press New York, NY.

Knapp, P., Raynor, D. K., Silcock, J., & Parkinson, B. (2011). Can user testing of a clinical trial patient information sheet make it fit-for-purpose?-a randomized controlled trial. *BMC medicine, 9*, 89.

Kreutzer, M. A., & Charlesworth, W. R. (1973). Infants' Reactions to Different Expressions of Emotions.

Lovato, L. C., Hill, K., Hertert, S., Hunninghake, D. B., & Probstfield, J. L. (1997). Recruitment for controlled clinical trials: literature summary and annotated bibliography. *Controlled clinical trials, 18*, 328-352.

Mello, R. A. (2002). Collocation analysis: A method for conceptualizing and understanding narrative data. *Qualitative research, 2*, 231-243.

Meyer, K. (n.d.). The Four Dimensions of Tone of VOice. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/tone-of-voice-dimensions/

MHRA. (2016). Best Practice Guidance on Patient Information Leaflets. MHRA. Retrieved from https://www.gov.uk/government/publications/best-practice-guidance-on-patient-information-leaflets

Mohammad, S. M., & Turney, P. D. (2013). *NRC emotion lexicon.* Tech. rep., NRC Technical Report.

Mohammad, S. M., & Yang, T. W. (2011). Tracking sentiment in mail: how genders differ on emotional axes. *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (acl-hlt 2011*, (pp. 70-79).

Moore, L., & Savage, J. (2002). Participant observation, informed consent and ethical approval. *Nurse Researcher, 9*, 58-69.

Moult, B., Franck, L. S., & Brady, H. (2004). Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expectations, 7*, 165-175.

MRC. (2016). Consent and Participant Information Sheet Preparation Guidance. NHS. Retrieved from http://www.hra-decisiontools.org.uk/consent/principles-general.html

National Institute for Health Research -NHS. (2014). *Patient and public involvement in health and social care research: A guide for researchers.* NHS.

NHS. (2017). Clinical trials. Retrieved from http://www.nhs.uk/Conditions/Clinical-trials/Pages/Introduction.aspx

Nicholls, S., Hankins, M., Hooley, C., & Smith, H. (2009). A survey of the quality and accuracy of information leaflets about skin cancer and sun-protective behaviour available from UK general practices and community pharmacies. *Journal of the European Academy of Dermatology and Venereology, 23*, 566-569.

O'Donnell, V., & Jowett, G. S. (1986). *Propaganda and persuasion.* Sage Publications.

OnlineOCR. (2016). *Online OCR -Convert PDF to Word or Image to Text*. Retrieved from OnlineOCR.net.

Pang, B., & Lee, L. (2004). Sentiment Polarity Dataset. *Proceedings of the 42nd annual meeting on Association for Computational Linguistics.* ACL.

Peterson, H. C. (1939). Propaganda for War, Norman. University of Oklahoma Press.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion, 1984*, 197-219.

Poplas-Susíc, T., Klemenc-Ketis, Z., Kersnik, J., & others. (2014). Usefulness of the patient information leaflet (PIL) and information on medicines from professionals: a patients' view. A qualitative study. *Zdravni{\v{s}}ki Vestnik, 83*, 368-375.

Prus, R. C. (1996). *Symbolic interaction and ethnographic research: Intersubjectivity and the study of human lived experience.* SUNY press.

PULSE. (n.d.). Designing Effective Leaflets. Pulse Creative Marketing. Retrieved from http://www.pulsecreativemarketing.co.uk/?featured\_posts=downloads-designing-effective-leaflet

Qu, Y., Shanahan, J., & Wiebe, J. (2004). Exploring attitude and affect in text: Theories and applications. *AAAI Spring Symposium) Technical report SS-04-07. AAAI Press, Menlo Park, CA.*

Reid, J. C., Klachko, D. M., Kardash, C. A., Robinson, R. D., Scholes, R., & Howard, D. (1995). Why people don't learn from diabetes literature: influence of text and reader characteristics. *Patient education and counseling, 25*, 31-38.

Reinert, C., Kremmler, L., Burock, S., Bogdahn, U., Wick, W., Gleiter, C. H., . . . Hau, P. (2014). Quantitative and qualitative analysis of study-related patient information sheets in randomised neuro-oncology phase III-trials. *European Journal of Cancer, 50*, 150-158.

Richard, M. A., Martin, S., Gouvernet, J., Folchetti, G., Bonerandi, J. J., & Grob, J. J. (1999). Humour and alarmism in melanoma prevention: a randomized controlled study of three types of information leaflet. *British Journal of Dermatology, 140*, 909-914.

Saldaña, J. (2015). *The coding manual for qualitative researchers.* Sage.

Sarle, W. S. (1990). Algorithms for clustering data. *Technometrics, 32*, 227-229.

School for Primary Care Research. (2017). *Patient and Public Involvement Payment.* NHS.

Schulz, K. P., Fan, J., Tang, C. Y., Newcorn, J. H., Buchsbaum, M. S., Cheung, A. M., & Halperin, J. M. (2004). Response inhibition in adolescents diagnosed with attention deficit hyperactivity disorder during childhood: an event-related FMRI study. *American Journal of Psychiatry, 161*, 1650-1657.

Starch, D. (1914). *Advertising: its principles, practice, and technique.* Scott, Foresman.

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science, 8*, 113-116.

The Writer. (n.d.). The Proof of the Pudding. TheWriter. Retrieved from http://www.thewriter.com/what-we-think/wisdom/the-writer-research/

Thomson, O. (1999). *Easily led: A history of propaganda.* Sutton Pub Limited.

Thomson, S. (1997). *The Folktale.* Berkeley: University of California Press.

Volcani, Y., & Fogel, D. (2006, 11). System and method for determining and controlling the impact of text. Google Patents.

Watson, J. B. (n.d.). *The psychology of advertising.*

Web Analytics Association. (2007). *Web Analytics Definitions*. Retrieved from DigitalAnalyticsAssociation.org: https://digitalanalyticsassociation.org/Files/PDF\_standards/WebAnalyticsDefinitionsVol1.pdf

Xiong, H., Wu, J., & Chen, J. (2009). K-means clustering versus validation measures: a data-distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*, 318-331.