UNIVERSITY OF SOUTHAMPTON

Employing Content Analysis & Crowd-sourcing to Revise Randomised Controlled Trials Patient Information Leaflets

by

MSc. Fernando Santos Sanchez

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Web and Internet Science Research Group
School of Electronics and Computer Science

June 2022

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, ELECTRONICS AND COMPUTER SCIENCE
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by MSc. Fernando Santos Sanchez

The poor readability of patient information leaflets (PILs) to help recruit people into research studies has been a serious concern for the Health Research Authority during the last 2 decades. Multiple independent studies have reported serious issues with almost all documents intended to inform patients or general audiences. The results of these studies have been considered by UK medical institutions and a series of guidelines developed to improve the quality and readability of PILs intended for inviting patients to randomised controlled trials (RCTs). Even with this focus and some improvements in patient leaflets, most of the current documents intended for this purpose are still considered too complex to be understood by public audiences.

Meanwhile, several techniques have addressed the issue of measuring text readability. This thesis analyses the utility of several of these techniques to identify the sentences that are too hard to understand when employed via a webtool to help PIL authors identify and correct PIL readability issues: a) readability indices that associate text characteristics with the US school grade needed to understand the document, b) the Cloze procedure (identifying specific words that are not understood by the participants by removing words and asking participants to complete the sentences), c) sentiment and content analysis to identify and associate comments from public participants reviewing PILs with specific sections of the documents, and d) online crowdsourcing to review and validate sentences that are too hard to be understood by public audiences.

The first study explored associations between PIL text characteristics and recruitment rates to the RCT to which they applied. The second studied feedback given by local public participants asked to revise PILs containing serious readability issues. The third study contained several sub-studies, each assessing the effects of the previously mentioned techniques on the identification, revision and validation of readability issues present in the PILs by an online crowd recruited using Mechanical Turk.

This thesis contributes to our knowledge about employing content analysis and online crowdsourcing techniques to help authors of PIL for RCTs to identify, revise and validate sentences that are too hard to be understood by public audiences.

Contents

List of Figures

List of Tables

Nomenclature

Crowdsourcing	Process of solving a problem by engaging a crowd or group of people.
EmotionLexicon	Weighted dictionary of associations between words and emotions or sentiments
MTurk	Amazon Mechanical Turk crowdsourcing platform
PIL	Patient Information Leaflet
PPI	Patient and Public Involvement
PragmaticTrial	Trials designed to assess the effectiveness of interventions in routine practice cond
RCT	Randomized Controlled Trial

Acknowledgements

I thank the NIHR Evaluation, Trials and Studies Coordinating Centre team and the main researchers from the Flupoc, Horizons and Truenth RCTs whose helpful contribution and the collaboration of the many public participants who permitted the realization of this work by providing the necessary data and feedback. 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.

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.

Chapter 1

Introduction

It is the intention of this research project to build a webtool to help create better information leaflets intended to inform potential participants in clinical trials. This platform employs the Amazon crowdsourcing platform Amazon (2017), content analysisGraham et al. (2017a); Krippendorff (2004) and readability indexes to support the quantitative assessment Patient Information Leaflet (PIL) readability Gray and Leary (1935); Gill et al. (2012) and its effect on recruitment, trial settings, participant feedback and the understanding of the PIL information.

This project first assessed the associations between the PILs readability, the presence of emotive content, the trial settings and the recruitment of participants in Chapter 4 where it was found readability correlates to the percentage of recruited participants, being a document for certain types of trials, the lexical complexity of the document, and the presence of words related to certain emotion classes.

As a second step, the feedback of public participants for PIL with known severe readability issues was studied. The results presented in Chapter 5 determined a sample of 30 participants could provide a good coverage of the readability issues by comparing the codes of the comments with the proposed topics in the EQIP scale Charvet-Berard et al. (2008) to ensure information quality; that participants with higher education levels failed to identify the readability issues as severe; and identified both readability and education were significant factors in predicting information understanding, but not qualitative assessments and subjective general comments were not.

The third step presented in Chapter 6 in our research included the development of a framework and webtool to collect, visualize, identify, revise and validate readability issues in PILs currently in use. Where it was found employing crowdsourcing could help improve the readability of specific sentences for a median of 2.5 grades, and that PIL authors considered essential to be guidance on how to correct the issues in addition to identifying and visualizing them.

1.1 Motivation

Clinical trials have become a corner stone Lovato et al. (1997) for identifying effective 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 (2017a). 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 and Savage (2002). Thus, to enforce ethical practice during recruitment, it is of great importance to ensure that patients considering participation are aware of the risks MRC (2016). Therefore, one of the core task for any clinical trialist is to develop Patient Information Leaflets (PILs) which are able to inform patients about essential trial features. The current clinical research process is based on the NHS proportionate approach to consent HRA (2017-01-17), which enables most PILs for randomised controlled trials (RCTs) to be designed by filling out template forms provided by the HRA and then reviewed by an ethics panel as part of the research submission process.

However, although this information is recognized as an essential part of any RCT by the HRA NHS (2017a), several independent studies in the last decade have consistently found that most PILs have serious deficits in informing patients, despite fulfilling the legal requirements and following NHS recommended guidelines and templates Reinert et al. (2014); Gillies et al. (2014); Poplas-Susíc et al. (2014); Knapp et al. (2011a); Nicholls et al. (2009) may not be fit to fulfill their purpose of supporting the consent process by helping ensure that all those who are invited to take part in a research study have been adequately informed MRC (2016). Several different approaches have sought to address these issues from employing quantitative content analysis of the PIL 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 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).

1.2 Thesis structure and justification

This thesis has been organized into three general parts describing the research process to consolidate the different facets of creating PILs that are easier to understand by members of the public by employing Web techniques. In the first part this research project focuses onto assessing the essential characteristics of PIL texts, determining their emotive composition and the feasibility of employing diverse text analysis techniques to assess their contents. The second part of the thesis explores the themes and composition of public comments given on PILs with severe readability issues from trials with poor recruitment rates. In the final part of the thesis an evaluation is made of 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.

1.2.1 Main Research Question

How can readability metrics, thematic or content analysis and crowdsourcing be used to help improve the readability of patient information leaflets from randomized controlled trials?

1.2.2 Main Objectives

The main objectives of this research are:

1. To explore the associations between the readability of PILs and recruitment rates to RCTs.
2. To characterise the feedback given by public commentators on PILs with serious readability issues.
3. To analyse the use of content analysis, readability metrics and crowdsourcing via a webtool to identify, revise and validate readability issues in PILs.
    (a) Analyzing the feedback from PIL authors on a webtool that employs content analysis and readability metrics to identify and visualize readability issues.
    (b) Analyzing the use of crowdsourcing to revise PIL sentences that were deemed too hard to be understood by public audiences.
    (c) Analyzing the use of crowdsourcing to validate proposed revisions to the original sentences.

Chapter 2

Literature review

2.1 Patient information leaflets

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 and most other countries of the world. This information is commonly presented as PILs, information sheets, online documents or videos that complement or enhance the explanations given by the trial recruiters. In the UK these documents are regulated by following the pre-set formats and guidelines on best practice for medical research set by the Health Research Authority (HRA). Under these guidelines the PILs must support 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 PIL documents NHS (2017a), several concerns have been raised about their quality in the last decade. The lack of a rigorous method for assessing the quality of written patient information, consent materials that are difficult to read Moult et al. (2004), inaccurate content Moult et al. (2004); Nicholls et al. (2009); Escudero-Carretero et al. (2013) and insufficient quality by most evaluated categories (e.g. text length, legibility, layout, visual structure) except ethical and legal requirements Reinert et al. (2014) were identified as high priority research topics by 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 capabilities of current Web model. These tasks are commonly called as text (data) mining and the set of techniques and models. These tasks are commonly called 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 computational challenges of working with unstructured data formats such as text have been recognized since the late 1950's Luhn (1958):

> ”...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 led to the creation of many techniques in areas like information retrieval, named entity recognition, disambiguation, co-reference, relationship and content analysis that could be of practical use when applied to the PILs. This project seeks to assess if a Web platform can make use of sentiment analysis, readability metrics, crowdsourcing and online recruitment to facilitate Public Involvement when revising information leaflets for potential participants of randomized controlled trials. This project also explore the effects of adding an information retrieval system (for previous public comments and writing guidelines) and content analysis reports as an enhancement 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 understanding documents and using sentiment analysis to detect the opinions and perceptions of the reviewers could also greatly enhance the feedback given to a trialist designing a new PIL. Thus, a Web platform was designed to:

- Collect public feedback on RCT PILs.
- Employ text analysis and readability metrics to objectively identify sentences that require higher reading skills than the average on general populations.
- Use the web platform to crowd-source the revision of PIL sentences with low readability.
- Employ the platform to validate the readability of these revisions.

We also provide analyses of the association between participant performance, sentence readability and participant reading skill level, and the effects of learning and fatigue on participants who revise the sentences.

2.2 PIL development process

It has been mentioned previously that providing the patient with information to make an informed decision is a fundamental part of trials in the UK NHS (2017a). This information generally includes PILs, information 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 (2017a); MRC (2016). 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 et al. (2004). As has been previously commented, the HRA guidelines encourage the researchers to employ heavily standardized forms and formats with only general advice given on how to describe the RCT. This advice consists 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 trialist community that the PILs must employ "everyday language" and explain complex words and clinical jargon but employ a "respectful tone" Charvet-Berard et al. (2008).

The HRA guidance documents "Applying a proportionate approach to the process of seeking consent" HRA (2017-01-17), "Consent and Participation Information Sheet Preparation Guidance" HRA (2014-03-03) and "Consent and Participant Information Sheet Preparation Guidance" MRC (2016) outlines the framework on how to design PILs for RCTs in accordance with UK legal requirements. These HRA guidelines focus on applying the principle of proportionality and creating more accessible participant information for clinical trials seeking participants. The main focus of this particular set of guidelines is clinical trials of medicinal products (CTIMPs), but it is also commonly applied to clinical trials of devices or other types of interventional/non-interventional research for Health Research NHS (2014).

The current proportionate approach for seeking consent tries to balance two divergent factors: that seeking informed consent is central to ethical research HRA (2017-01-17) and that seeking consent has become a rigid perfunctory procedure Afolabi et al. (2014); Hansson (1998); Ploug and Holm (2013); Tobias and Souhami (1993) with information sheets that are too complex to help potential participants Roberts et al. (2011). This implies that the need to give potential research participants the necessary information to help them decide about participating has been overtaken by documents whose principal function has become to protect researchers and sponsors from litigation by describing every minor detail Varnhagen et al. (2005); O'Neil et al. (2003). Thus, the current proportionate approach seeks to implement procedures that correspond to the balance of risk and benefits to avoid lengthy and complex information leaflets. Creating user-friendly information leaflets that contain succinct, relevant, truthful information is the ultimate goal of these guidelines HRA (2017-01-17). Therefore, the closer the research is to current clinical practice, the less detail is needed in the information leaflet, suggesting that in many accounts it will be the verbal exchange during the discussion with the potential participant that will be crucial in facilitating the decision HRA (2017-01-17). The HRA current guidelines are based on 14 principles drawn from the Medicines for Human Use (Clinical Trials) regulations HRA (2017-01-17); MRC (2016).

1. The rights, safety and well-being of the trial subjects shall prevail over the interests of science and society.

2. Each individual involved in conducting a trial shall be qualified by education, training and experience to perform his tasks.

3. Clinical trials shall be scientifically sound and guided by ethical principles in all their aspects.

4. The necessary procedures to secure the quality of every aspect of the trial shall be complied with.

5. The available non-clinical and clinical information on an investigational medicinal product shall be adequate to support the proposed clinical trial.

6. Clinical trials shall be conducted in accordance with the principles of the Declaration of Helsinki.

7. The protocol shall provide for the definition of inclusion and exclusion subjects participating in a clinical trial, monitoring and publication policy.

8. The investigator and sponsor shall consider all relevant guidance with respect to commencing and conducting a clinical trial.

9. All clinical information shall be recorded, handled and stored in such a way that it can be accurately reported, interpreted and verified, while the confidentiality of records of the trial subjects remains protected.

10. Before the trial is initiated, foreseeable risks and inconveniences have been weighed against the anticipated benefit for the individual trial subject and other present and future patients. A trial should be initiated and continued only if the anticipated benefits justify the risks.

11. The medical care given to, and medical decisions made on behalf of, subjects shall always be the responsibility of an appropriately qualified doctor or, when appropriate, of a qualified dentist.

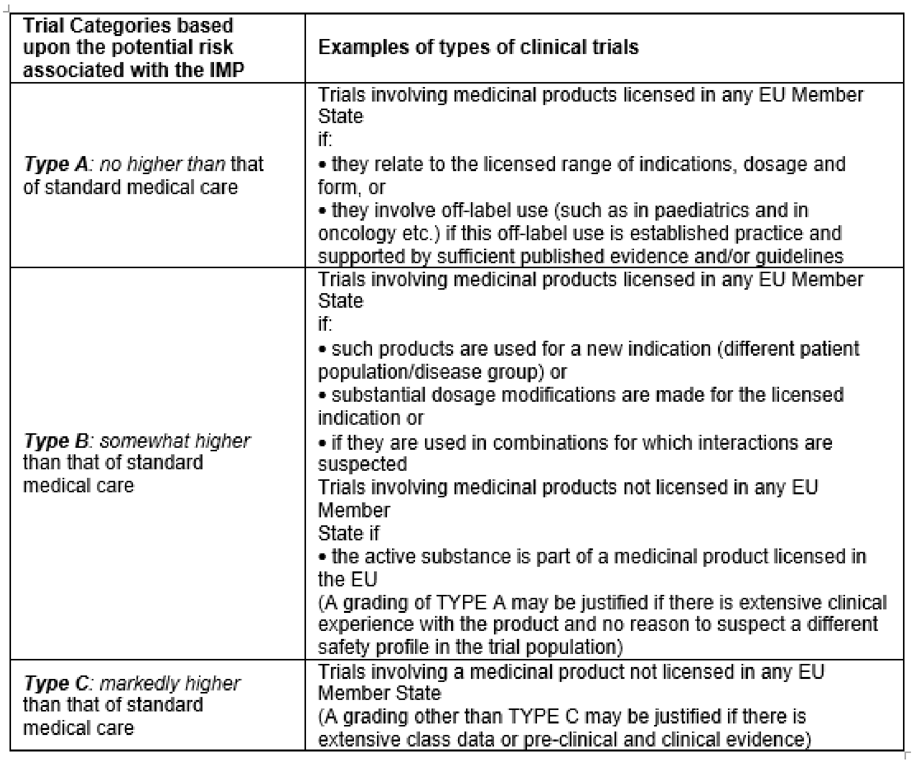
12. A trial shall be initiated only if an ethics committee and the licensing authority comes to the conclusion that the anticipated therapeutic and public health benefits justify the risks and may be continued only if compliance with this requirement is permanently monitored.

13. The rights of each subject to physical and mental integrity, to privacy and to the protection of the data concerning him in accordance with the Data Protection Act are safeguarded.

14. Provision has been made for insurance or indemnity to cover the liability of the investigator and sponsor which may arise in relation to the clinical trial.

These principles and common law require that participants “be informed, in broad terms, of the nature and purpose of the research and the material risks, and benefits and reasonable alternatives” HRA (2017-01-17). Therefore, the core information about a trial should be provided in a succinct form, paying attention to the way it is conveyed, using language that most people can understand and considering the layout and format to aid the explanation.

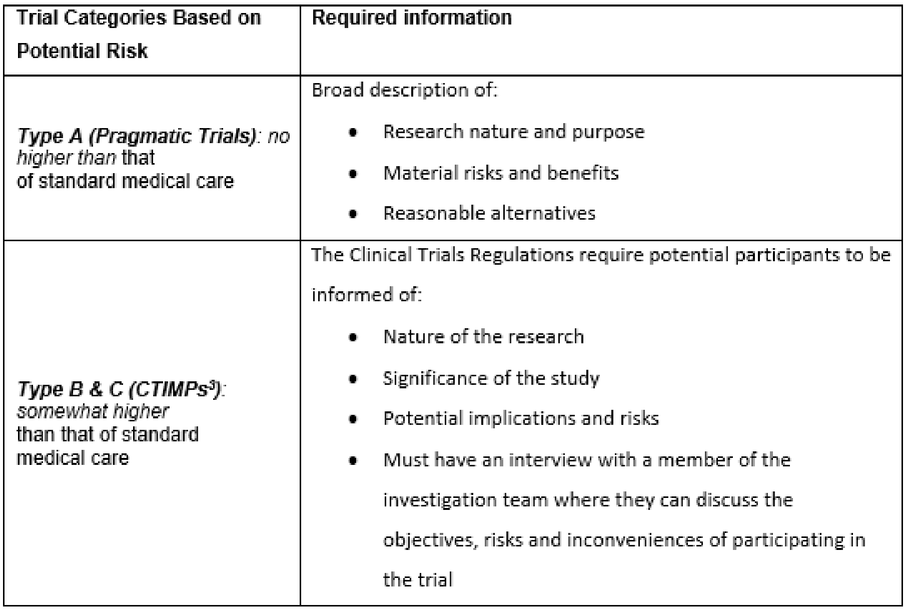
These has lead the HRA to consider that the amount of information to be provided to participants outside the core information (research nature, significance, implications and risks) when seeking their participation must vary in accordance with the balance between risk and benefits of the research e.g. practical information about the trial (timings, payment of travel expenses, etc.) would only be needed if it has implications for the participant decision to join the trial (need for abstinence, significant drug interactions, etc.). The MHRA categorises three levels of trial risk, with pragmatic trials considered a special subset within these guidelines as they generally do not involve additional risk to those inherent in current care practices. In pragmatic trials it should often be possible to simplify the necessary information into a single, short participant information sheet. Pragmatic trials, also known as ‘simple trials’, ‘comparative effectiveness trials’, ‘non-interventional trials’ or ‘low-intervention trials’, are defined as trials that do not involve interventions beyond the normal care of the patient, rather they focus on comparing the effects of accepted/licensed interventions or therapies in current clinical practice, see Figure 2.1.



Trial Categories based upon the potential risk associated with the IMP	**Examples of types of clinical trials**
***Type A**: no higher than* that of standard medical care	Trials involving medicinal products licensed in any EU Member State if: • they relate to the licensed range of indications, dosage and form, or • they involve off-label use (such as in paediatrics and in oncology etc.) if this off-label use is established practice and supported by sufficient published evidence and/or guidelines
***Type B**: somewhat higher* than that of standard medical care	Trials involving medicinal products licensed in any EU Member State if: • such products are used for a new indication (different patient population/disease group) or • substantial dosage modifications are made for the licensed indication or • if they are used in combinations for which interactions are suspected Trials involving medicinal products not licensed in any EU Member State if • the active substance is part of a medicinal product licensed in the EU (A grading of TYPE A may be justified if there is extensive clinical experience with the product and no reason to suspect a different safety profile in the trial population)
***Type C**: markedly higher* than that of standard medical care	Trials involving a medicinal product not licensed in any EU Member State (A grading other than TYPE C may be justified if there is extensive class data or pre-clinical and clinical evidence)

FIGURE 2.1: Clinical trials categories based upon the potential risk to the patient. -HRA (2017-01-17)

Pragmatic trials involving non-drug interventions only need to comply with the common law, but research involving medicines also needs to comply with "The Medicines for Human Use (Clinical Trials) Regulations" Parliament (2004) referred as Clinical Trial Regulations. The Clinical Trial Regulations also apply to pragmatic trials where the research protocol is used to decided what drug is given to the patients instead of their doctors or other healthcare professional as part of their clinical care, Figure 2.2.



Trial Categories Based on Potential Risk	Required information
***Type A (Pragmatic Trials)**: no higher than* that of standard medical care	Broad description of: • Research nature and purpose • Material risks and benefits • Reasonable alternatives
***Type B & C (CTIMPs[3])**: somewhat higher* than that of standard medical care	The Clinical Trials Regulations require potential participants to be informed of: • Nature of the research • Significance of the study • Potential implications and risks • Must have an interview with a member of the investigation team where they can discuss the objectives, risks and inconveniences of participating in the trial

Figure 2.2: Required information to be given to the patient based on RCT risk category. -HRA (2017-01-17)

The HRA guidelines include a PIL template for RCTs HRA (2017-01-17) to be used and adapted for pragmatic trials and Type B and C CTIMPs, which is also commonly employed as a reference for other research studies (presented in section 2.3.1). To complement these principles, the HRA “Consent and Participant Information Sheet Preparation Guidance” HRA (2014-03-03) provides further guidance on how to create good information for potential participants, by:

“1. Taking notice that the information required to enable potential participants’ decision will vary in accordance with the nature and burden of the research.

2. Creating PILs as simple and short as possible while including all necessary information to enable the participant decision.

3. Setting the importance of your study, designing a good title that provides a concise summary of the study with words your participants can understand.

4. Employing an invitational style, create a PIL that is a polite invitation to participate, setting potential advantages, risks and alternatives.

5. Do not employ passive voice.

6. Employing plain English and avoiding clinical terminology (jargon) when possible.

   (a) Remember your audience

   (b) Use short words and sentences

   (c) Use lay language and familiar words to your audience

   (d) The language should not be more difficult than medicine leaflets or tabloid newspapers

   (e) Participants should understand the PIL in the first reading

   (f) All potential participants should understand your PIL

   (g) Limit sentences to no more of 20 words

   (h) Do not include more than one idea per sentence. If the next sentence does not follow the previous one, start a new paragraph

   (i) Avoid obscure or commonly misunderstood words (dual or nuanced meanings e.g. drugs and diet)

   (j) Avoid more than two hard words in a sentence unless you are explaining a term and consider employing acronyms for repeated use. A hard word is a word that is a technicism, jargon, uncommon, long or with many syllables.

7. Use a format that support understanding

   (a) Use short heading that stand out

   (b) A question-answer format is effective

   (c) Use large type size (16 pts) if you are recruiting elderly subjects

   (d) Avoid unbroken sections of text or long lists

   (e) Use bullet points for lists

   (f) Avoid justified text

   (g) Use bold lower case for emphasis

   (h) Consider the use of multimedia to support the consent process (CDs, DVDs, etc.)

8. Consider the use of diagrams to facilitate the explanation and discussion with the participant

9. Consider the participant perspective, address issues that may be very important to the participants' decision (e.g. Will I have to take time off to take part? How many times will I need to attend?)

10. Be clear about expected risks and benefits

11. If you are recruiting two or more groups of participants, consider creating different PILs to address their particular concerns

12. Test your PIL with and appropriate group of people (Patient or Public groups), you do not need NHS Research Ethics Committee (REC) approval to test your consent documents"–HRA (2014-03-03)

Additional guidance is given in the document HRA (2014-03-03) for Adults who are unable to consent by themselves, children and young people and emergency research. These topics fall outside the scope of this research and thus are omitted from further consideration.

2.3 PIL quality issues

Most recent research on PILs has focused on determining their quality or developing an objective method of measuring their quality, in response to Moult Moult et al. (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 et al. (2004) Nicholls et al. (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 et al. (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 (2017a). 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 et al. (2011a) Escudero-Carretero et al. (2013) Gillies et al. (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 et al. (2011a) to the employment of specialized formulas and instruments like the Flesch-Kincaid Gillies et al. (2014) or Flesch-Formel Reinert et al. (2014) coefficients and the SMOG/INFLESZ scores Escudero-Carretero et al. (2013). In addition, Knapp Knapp et al. (2011a) carried out qualitative work to measure reading times, interest in the topics, and comprehension of the topics finding the original documents may not have enabled valid consent as only 15% of the readers understood all the aspects in the PILs. 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 et al. (2009) Gillies et al. (2014) Reinert et al. (2014). Reinert's study on neuro-oncology phase III trial PILs Reinert et al. (2014) determined that five (56%) 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 finding that all documents were of insufficient quality in all categories except the fulfilment of ethical and legal considerations 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 et al. (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 31 PILs for skin cancer found that all but one PIL required education above primary level. A qualitative study on drug PILs Poplas-Susíc et al. (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 et al. (2011a), 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 et al. (2011a).

Reinert's results show that "All patient informed consent documents (9 PILs) 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 systematic review of 14 instruments that found the PILs provided for trials on UK Clinical Trial Unit websites did not support good quality decision-making, the existence of variability in the conceptualization, development and domain coverage of the measures for assessing informed consent, a narrow focus on the considerations of decision making, and a lack of identification of key domains to assess informed consent Gillies et al. (2014). A summary of findings on PIL issues presented in this section is provided in Table 2.1.

Table 2.1: PIL issues.

Study Title	PIL Aspect Studied	Identified issues	Reference	Impact
Ensuring Quality Information for Patients: development and preliminary validation of a new instrument to improve the quality of written health care information	Design and validation of an instrument to assess the quality of written health information		B. Moult, L.S. Franck, and H. Brady. 2004.	The EQIP instrument was demonstrated to have significant correlation with DISCERN, good validity, reliability and utility
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	Survey assessment of information quality and accuracy using readability indexes (SMOG) and the EQIP guidelines	31PILs were analysed, all PILs were required higher reading skill than the recommended range for health education, 17% of the documents contained mayor inaccuracy	S. Nicholls, M. Hankins, C. Hooley, and H. Smith. 2009.	Accuracy of the PILs varied greatly, less than half were judge to be fully accurate. Current PILs do not satisfy the needs of patients with low literacy

Continued on next page

Table 2.1 – continued from previous page

Study Title	PIL Aspect Studied	Results	Reference	Impact
Usefulness of the patient information leaflet (PIL) and information on medicines from professionals: a patients' view. A qualitative study	Analysis of PIL usefulness from the patient perspective	Patients read selectively and found language to be too scientific. Current PIL offer partial information, contain legibility issues and do not enable comprehensive information.	T. Poplas-Sus´ıc, Z. Klemenc-Ketis, and J. Kersnik. 2014.	Patients were most interested in side effects and contraindications; considered a family physician to be the most trustworthy source of information and consider pharmacist could play a more active role in educating patients
				Continued on next page

Table 2.1 – continued from previous page

Study Title	PIL Aspect Studied	Results	Reference	Impact
Why people don't learn from diabetes literature: influence of text and reader characteristics	Study focus on identifying text and reader characteristics that impede learning	26 adult patients assessed prior knowledge of diabetes and reading skill. Participants could only recall on average 8 of 108 ideas present on a commonly use diabetes pamplet immediately after reading it and the lack of clarity and organization in the documents can hinder comprehension	J.C. Reid, D.M. Klachko, C.A. Kardash, R.D. Robinson, R. Scholes, and D. Howard. 1995.	Readers seldom monitor their comprehension, the topics identified as important by patients and physicians differ

Continued on next page

Table 2.1 – continued from previous page

Study Title	PIL Aspect Studied	Results	Reference	Impact
Can user testing of a clinical trial patient information sheet make it fit-for-purpose?	Controlled design to assess the effect of user testing to test and improve readability	The performance of 50 participants of the target group (myeloid leukemia AML16) was assessed on the original PIL. In a second phase 123 participants assessed a revised version. The original PIL may not have enabled valid consent, only 15% of the participants reading the original PIL understood all concepts compared against the 68% on the revised document	P. Knapp, D.K. Raynor, J. Silcock, and B. Parkinson. 2011.	User testing significantly improve the readability of the document, the original PIL may not be fit for purpose

Continued on next page

Table 2.1 – continued from previous page

Study Title	PIL Aspect Studied	Results	Reference	Impact
Patient information leaflets (PILs) for UK	Assessment of RCT PILs by employing an evaluation tool based on the standards for supporting decision making	Majority of the 20 PIL analysed scored poorly (below 50%), the score was found to be associated to the word count of the PILs but not to their readability scores.	K. Gillies, W. Huang, Z. Skea, J. Brehaut, and S. Cotton. 2014.	Most documents have issues presenting probabilities, clarifying and expressing values, and presenting structured guidance in deliberation and communication
Quantitative and qualitative analysis of study-related patient information sheets in randomised neuro-oncology phase	Analysis of PIL quality for phase III studies (neuro-oncological) based on text length, layout, readability, application of ethical and legal considerations, scientific evidence, and social aspects	All the documents (9 PILs) were found to have poor quality in all areas other than fulfiling the ethical and legal requirements. Graduate levels were required to read and understand 5 of the 9 documents	C. Reinert, L. Kremmler, S. Burock, U. Bogdahn, W. Wick, C.H. Gleiter, and P. Hau. 2014.	Restructuring document layout into patient friendly design is necessary, standardize components focusing on the patient knowledge, mode of treatment and time of use should be incorporated.

2.3.1 HRA guidelines to seeking consent

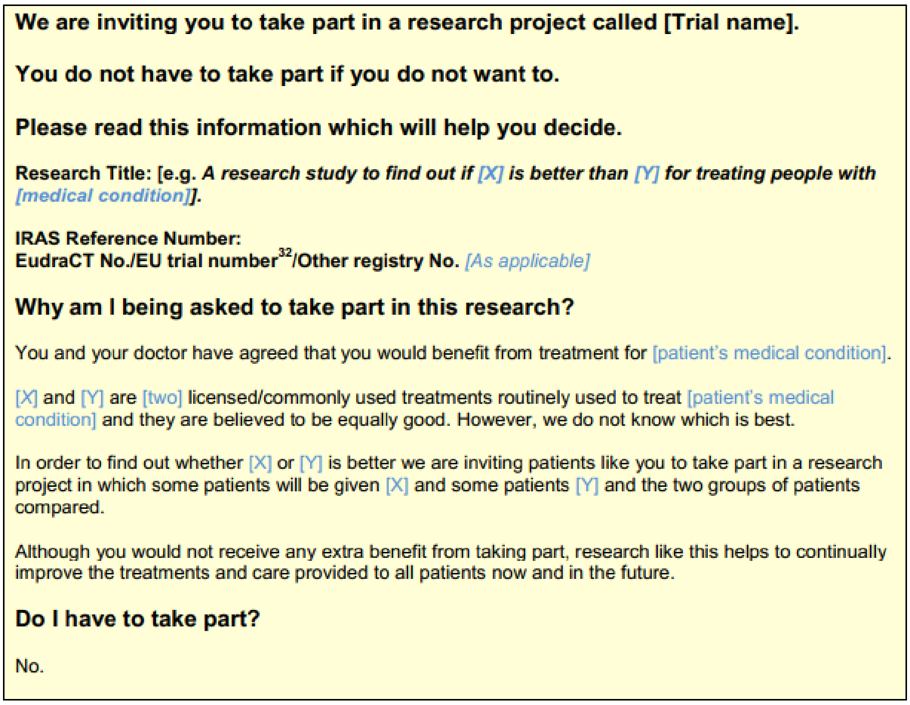
The current approach of seeking consent for clinical research is based on the ethical and legal principles coded in Health Research Authority (HRA) guidelines and framework HRA (2017-01-17, 2020). The HRA reviews and approves all clinical research in the UK to ensure it is ethically designed and a high-quality research standard is maintained. The guidance provided to clinical researchers by the HRA focus on:

- Ethical and legal principles of consent
- The application of these principles to designing Patient Information Leaflets (PILs) and consent forms
- Recommended content of PILs and consent forms
- Recommended style and design of PILs and consent forms

The HRA proportionate approach to the process of seeking main objective is to create better information for potential participants by adjusting the level of detail that must be included based on the balance between the benefits and risks of the trial over current care practice for the patient. The HRA has determined that for consent to be legal and ethical it must be:

1. Given by a person with capacity
2. Voluntary, with no undue influence
3. Given by a person who has been adequately informed
4. A fair choice

The HRA guidelines recognize three levels of risks for clinical trials, type A when there is no additional risk than the normal care to the patient, type B when participation inherently includes additional risks to the patient than those expected from current care, and type C when there is significant risk to the patient. The HRA provides a PIL template (Figure 2.3,Figure 2.4,Figure 2.5) for medicinal clinical trials that can be adjusted to the requirements of type A, B, and C trials HRA (2017-01-17). This template is also recommended for other types of clinical research HRA (2017-01-17).



We are inviting you to take part in a research project called [Trial name].

You do not have to take part if you do not want to.

Please read this information which will help you decide.

Research Title: [e.g. *A research study to find out if [X] is better than [Y] for treating people with [medical condition]*].

IRAS Reference Number:
EudraCT No./EU trial number[32]/Other registry No. *[As applicable]*

Why am I being asked to take part in this research?

You and your doctor have agreed that you would benefit from treatment for [patient's medical condition].

[X] and [Y] are [two] licensed/commonly used treatments routinely used to treat [patient's medical condition] and they are believed to be equally good. However, we do not know which is best.

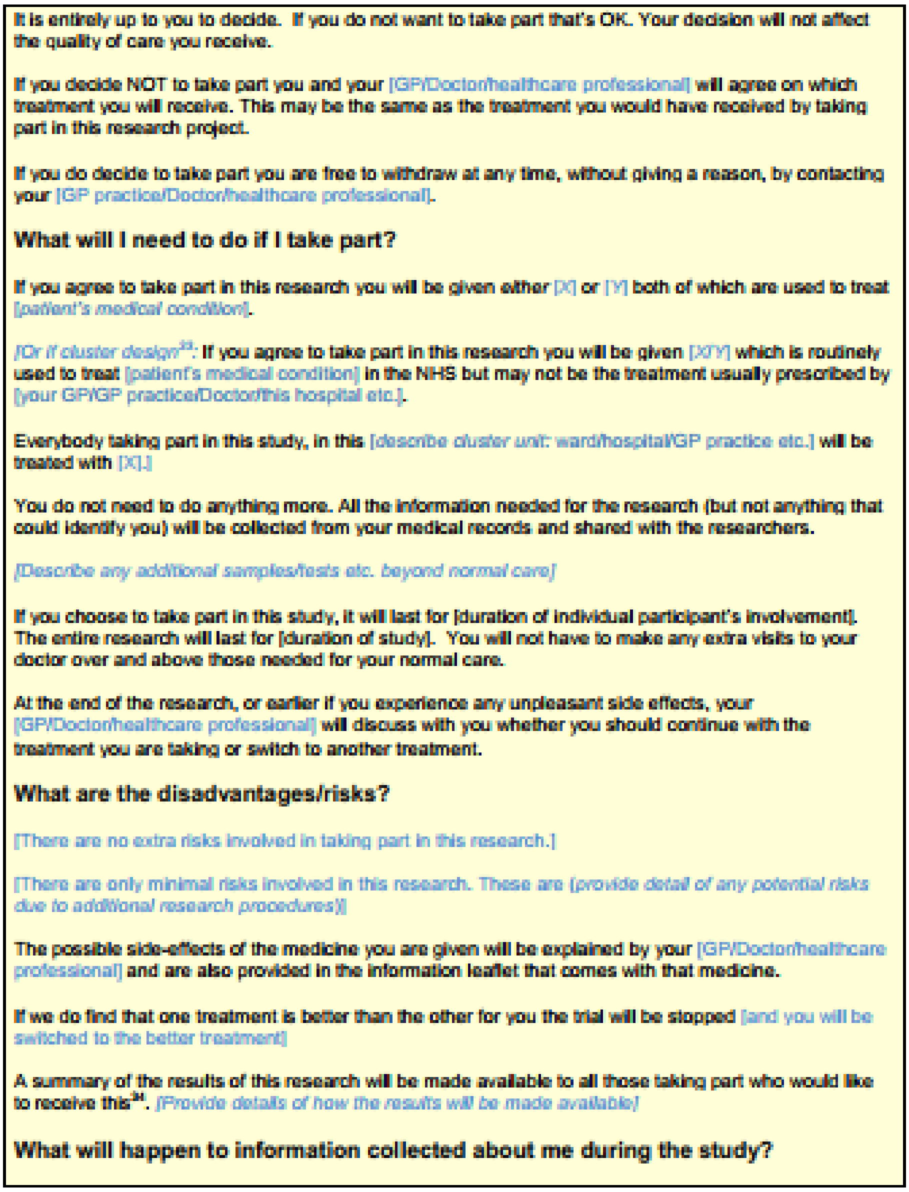
In order to find out whether [X] or [Y] is better we are inviting patients like you to take part in a research project in which some patients will be given [X] and some patients [Y] and the two groups of patients compared.

Although you would not receive any extra benefit from taking part, research like this helps to continually improve the treatments and care provided to all patients now and in the future.

Do I have to take part?

No.

Figure 2.3: HRA PIL Template 2017 Part1.



It is entirely up to you to decide. If you do not want to take part that's OK. Your decision will not affect the quality of care you receive.

If you decide NOT to take part you and your [GP/Doctor/healthcare professional] will agree on which treatment you will receive. This may be the same as the treatment you would have received by taking part in this research project.

If you do decide to take part you are free to withdraw at any time, without giving a reason, by contacting your [GP practice/Doctor/healthcare professional].

What will I need to do if I take part?

If you agree to take part in this research you will be given *either* [X] or [Y] both of which are used to treat *[patient's medical condition]*.

[Or if cluster design[33]: If you agree to take part in this research you will be given [X/Y] which is routinely used to treat [patient's medical condition] in the NHS but may not be the treatment usually prescribed by [your GP/GP practice/Doctor/this hospital etc.].

Everybody taking part in this study, in this [*describe cluster unit:* ward/hospital/GP practice etc.] will be treated with [X].]

You do not need to do anything more. All the information needed for the research (but not anything that could identify you) will be collected from your medical records and shared with the researchers.

[Describe any additional samples/tests etc. beyond normal care]

If you choose to take part in this study, it will last for [duration of individual participant's involvement]. The entire research will last for [duration of study]. You will not have to make any extra visits to your doctor over and above those needed for your normal care.

At the end of the research, or earlier if you experience any unpleasant side effects, your [GP/Doctor/healthcare professional] will discuss with you whether you should continue with the treatment you are taking or switch to another treatment.

What are the disadvantages/risks?

[There are no extra risks involved in taking part in this research.]

[There are only minimal risks involved in this research. These are (*provide detail of any potential risks due to additional research procedures*)]

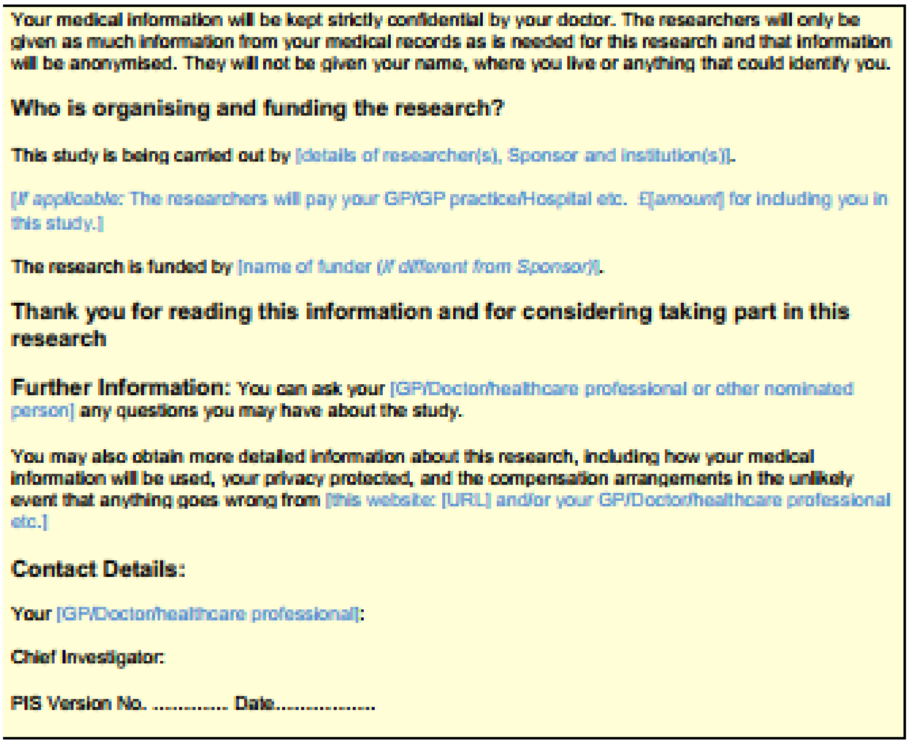
The possible side-effects of the medicine you are given will be explained by your [GP/Doctor/healthcare professional] and are also provided in the information leaflet that comes with that medicine.

If we do find that one treatment is better than the other for you the trial will be stopped [and you will be switched to the better treatment]

A summary of the results of this research will be made available to all those taking part who would like to receive this[34]. *[Provide details of how the results will be made available]*

What will happen to information collected about me during the study?

Figure 2.4: HRA PIL Template 2017 Part2.



Your medical information will be kept strictly confidential by your doctor. The researchers will only be given as much information from your medical records as is needed for this research and that information will be anonymised. They will not be given your name, where you live or anything that could identify you.

Who is organising and funding the research?

This study is being carried out by [details of researcher(s), Sponsor and Institution(s)].

[If applicable: The researchers will pay your GP/GP practice/Hospital etc. £[amount] for including you in this study.]

The research is funded by [name of funder *(if different from Sponsor)*].

Thank you for reading this information and for considering taking part in this research

Further Information: You can ask your [GP/Doctor/healthcare professional or other nominated person] any questions you may have about the study.

You may also obtain more detailed information about this research, including how your medical information will be used, your privacy protected, and the compensation arrangements in the unlikely event that anything goes wrong from [this website: [URL] and/or your GP/Doctor/healthcare professional etc.]

Contact Details:

Your [GP/Doctor/healthcare professional]:

Chief Investigator:

PIS Version No. .... Date.................

Figure 2.5: HRA PIL Template 2017 Part3.

In accordance with the proportionate approach the amount of detail included in each of the template sections must correspond to the level of risk the participant may face. The form provided by the HRA addresses the requirement of the common law and the UK Clinical Trial Regulations HRA (2017-01-17); Parliament (2004):

- Describe the nature and purpose of the research
- The significance of the study
- The potential implications, risks and benefits
- Reasonable alternatives

As a set of 6 questions directed to the participant:

- Why am I being asked to take part in this research?
- Do I have to take part?
- What will I need to do if I take part?
- What are the disadvantages/risks?
- What will happen to information collected about me during the study?

- Who is organizing and funding the research?

It also provides a template for practical information about the trial:

- Research title
- IRAS reference number
- Other registry number
- Lead researcher
- Research funder
- Contact details
- Link for further information
- PIL version
- Date

The amount of detail on each PIL would depend on its risk classification, type A research studies (pragmatic trials) would be able to have greater simplification often covering all relevant topics in a single page, while type B and C trials need to provide additional practical information when this information may have a direct impact on the potential participant decision (e.g. need for abstinence). The HRA proportional approach to consent has facilitated the creation of less cluttered and overwhelming PILs for clinical trials with low-risk for the patients. On the other hand, it has also induced a perfunctory adherence by clinical researchers on pragmatic trials Afolabi et al. (2014); Hansson (1998); Ploug and Holm (2013); Tobias and Souhami (1993) and has sown the idea that is the verbal discussion with the participants that is crucial to the potential participant decision among the assessing bodies HRA (2017-01-17). This has brought many cases were PILs for pragmatic trials are approved even when containing readability issues, as not enough focus is given to assess their ”information quality”. The next section discusses how information quality can be assessed.

2.4 The assessment of information quality

2.4.1 Patient and public involvement groups

The HRA “Consent and Patient Information Sheet Guidance” HRA (2017-01-17) guidance encourages the clinical researchers to test their PIL with appropriate Patient or Public groups, stating that doing so can help ensure that:

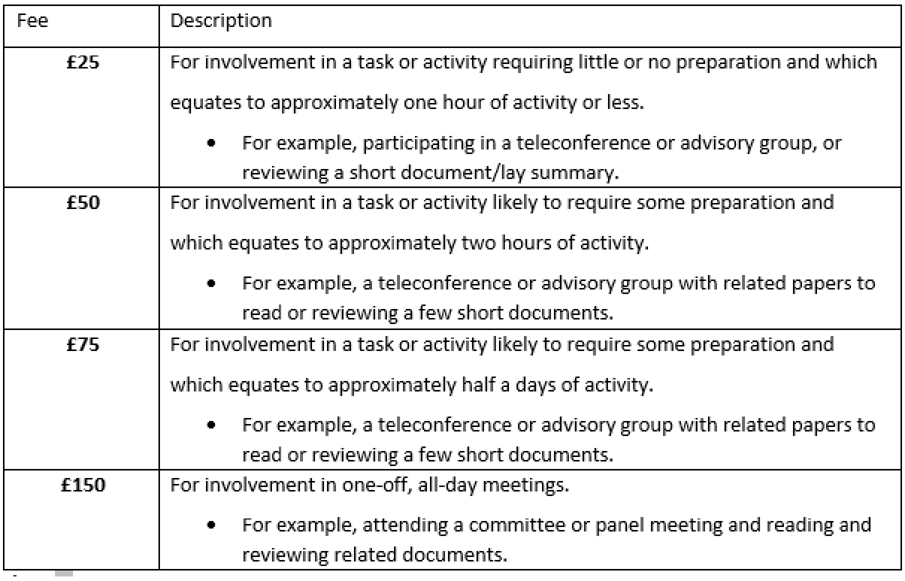
- The document employs appropriate language
- The style and format aids understanding
- The document covers the relevant risks and benefits to the potential participants.

While there is no need to obtain NHS Research Ethics Committee (REC) approval to approach members of the public to test a PIL, this may not be as simple a task as it appears. The current guidance on involving the public in clinical research is described in “Patient and Public Involvement in Health and Social Care Research” for Health Research NHS (2014) and in the INVOLVE website for Health Research NHS (2018).

In accordance with INVOLVE definition of public involvement, this is “research that is carried out with or by members of the public rather than to, about or for them” for Health Research NHS (2018). The term ”public” in this definition can include patients, potential patients, carers and people who use health and social care services, but seeks to differentiate public involvement from other activities. Under the Involve definition ”public involvement” is not raising awareness of research, sharing knowledge or engaging in dialog with the public. It also does not refer to the recruitment of patients or members of the public as participants in research. This means, that while the researchers may engage a patient or public involvement (PPI) group to revise their PIL, assessing the participants’ understanding of the PIL information falls outside the current definition provided by INVOLVE. In accordance with the NIHR guidelines for PPI for Health Research NHS (2014) the institute may ask to “applications that are technically excellent” to engage a PPI group before granting funding to the research, it also states that the main focus of the NIHR PPI activities since 2006 has been to support public involvement in the commissioning process for national research programs and that it expects all applications to be equally committed to PPI.

Another major concern when engaging PPI groups is the topic of representation. With PPI groups tending to be small, commonly less than 5 participants, it has been brought as a concern Martin (2008) that they lack representativeness of the intended population and where participants with odd views can have a disproportionate impact in the document because they are the only participants hea (2020).

The “Patient and Public Involvement Payment” guidance NHS (2017b) for PPI groups is to offer a contributor a payment or “involvement fee” and reasonable travel expenses. It defines public contributors as members of the public who are being asked to provide a public perspective and are not undertaking the task as part of their full time employment. The recommended fees for PPI contributors range from £25 per person per hour to £150 per day based on the complexity of the required tasks, Figure 2.6. When this cost is added to the proportionality principle of seeking consent, it leads most pragmatic trialists to the conclusion that engaging a PPI group is not a viable method to revise their PILs.



Fee	Description
£25	For involvement in a task or activity requiring little or no preparation and which equates to approximately one hour of activity or less. • For example, participating in a teleconference or advisory group, or reviewing a short document/lay summary.
£50	For involvement in a task or activity likely to require some preparation and which equates to approximately two hours of activity. • For example, a teleconference or advisory group with related papers to read or reviewing a few short documents.
£75	For involvement in a task or activity likely to require some preparation and which equates to approximately half a days of activity. • For example, a teleconference or advisory group with related papers to read or reviewing a few short documents.
£150	For involvement in one-off, all-day meetings. • For example, attending a committee or panel meeting and reading and reviewing related documents.

Figure 2.6: INVOLVE PPI Recommended Fees 2017. –NHS (2017c)

2.4.2 Crowd-sourcing

Crowd-sourcing can be defined as a model to create goods and services by incorporating ideas and resources from online groups of participants Schenk et al. (2009). The term became popular in 2006 as a derivative of ”crowd outsourcing” and currently is mostly applied to the recruitment of individuals through Web platforms to solve micro-tasks, which contribute to the development of solutions and services to organizations, researchers and public entities Howe (2006). Crowd-sourcing models are commonly classified by their approach as Brabham (2013a):

“1. Knowledge discovery and management

2. Distributed intelligence tasking

3. Broadcast searching of solutions for ideation problems with objective solutions

4. Peer vetted solution creation for ideation problems with subjective solutions or dependent on public support”. –Brabham (2013a)

Controversial topics have been brought on the use of crowd-sourcing in recent years: Borst et al. (2018); Aitamurto et al. (2011); Ross et al. (2010); Graham et al. (2017a,b); Budak et al. (2016); Brabham (2013b); Kleemann et al. (2008)

1. Impact of crowd-sourcing on product quality

2. Entrepreneurs contribute less capital themselves

3. Increased number of funded ideas
4. The value and impact of the work received from the crowd
5. The ethical implications of low wages paid to crowd-workers
6. Trustworthiness and informed decision making

Most of the controversy of these topics originates from the low payments offered to the participants when compared to the minimum wage regulations. There is a point on medical research that participants should not be induced into studies by economic gain, but that their remuneration should correspond to the expected risk they may encounter. All tasks in expected of participants in this research studies do not present a risk for the participants and thus it has been considered that the utilization of the average remuneration, $1 usd per participant per task, was deemed appropriate for tasks expected to be completed in less than 20 min.

This research project employed the Amazon Mechanical Turk platform to publish Human Intelligence Tasks (HITs) for three of the primary studies. First, the participants were employed in distributed intelligence tasking to read and identify readability issues on PIL texts. As a following step the participants were asked to revise sentences that were deemed to be too hard to understand by public audiences. Finally, a peer vetted crowdsourcing model was incorporated into the system to assess the validity of the proposed solutions.

Similar approaches have been used to crowdsource the translation of text, were it has been demonstrated that it is possible to obtain high quality translations from non-professional translators Zaidan and Callison-Burch (2011), and it has been compared to other automated approaches to translate text like machine learning were it was found to have high volume, quick translation output and low cost Anastasiou and Gupta (2011).

In addition, independent research has demonstrated that crowdsourcing in Amazon Mechanical Turk (MTurk) can help reach large and diverse samples at low cost Gosling and Mason (2015); avoid payment hassles that may decentivize participation Mason and Suri (2012); and attracts enough users to fullfil the researchers needs in clinical research Shapiro et al. (2013), and that while MTurk is not representative of the population at large it is still more diverse than samples commonly used in clinical research Paolacci et al. (2010).

2.4.3 Content analysis

Generally, content analysis is research using the categorization and classification of speech, written text, interviews, images, or other forms of communication. In its beginnings, using the first newspapers at the end of the 19th century, analysis was done manually by measuring the number of columns given to a specific subject. The approach can also be traced back to a university student studying patterns in Shakespeare's literature in 1893 Sumpter (2001).With the rise of common computing facilities like PCs, computer-based methods of analysis are growing in popularity Pfeiffer et al. (1997); Grimmer and Stewart (2013); Yi et al. (2003).

Content analysis methods all involve the systematic reading or measurement of texts or documents which are assigned labels (sometimes called codes) to indicate the presence of interesting, meaningful pieces of content Hodder (1994) Bell et al. (2018). The analysis can be carried out by humans or automated through computer systems. Social scientists use content analysis to examine patterns in communication in a replicable and systematic manner. The six essential questions of content analysis Krippendorff (2004) are defined as:

1. Which data are analysed?
2. How are the data defined?
3. From what population are data drawn?
4. What is the relevant context?
5. What are the boundaries of the analysis?
6. What is to be measured?

Quantitative content analysis highlights frequency counts and objective analysis of these coded frequencies Kracauer (1952). Typically, quantitative content analysis is deductive, beginning with a framed hypothesis with coding decided on before the analysis begins. These coding categories are strictly relevant to the researcher's hypothesis. Quantitative analysis can also take an inductive approach. White and Marsh (2006) in which the codes are based on the text, with no prior hypothesis.

Siegfried Kracauer provides a critique of quantitative text analysis, asserting that it oversimplifies complex communications in order to be more reliable. Qualitative text analysis, on the other hand, deals with the intricacies of latent interpretations, whereas quantitative has a focus on manifest meanings. Kracauer also acknowledges an "overlap" of qualitative and quantitative content analysis Kracauer (1952). When patterns are looked at more closely in qualitative analysis and based on the latent meanings that the researcher may find, the course of the research could be changed. This method is inductive and begins with open research questions, as opposed to a hypothesis White and Marsh (2006).

A study found that human coders were able to evaluate a broader range and make inferences based on latent meanings Conway (2006). Robert Weber notes: *"To make* *valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way”* Weber (1990). The validity, inter-coder reliability and intra-coder reliability have been subject to prolonged, intense methodological research efforts Krippendorff (2004). Neuendorf suggests that when human coders are used in content analysis at least two independent coders should be used. The reliability of human coding is often measured using a statistical measure of inter-coder reliability or ”the amount of agreement or correspondence among two or more coders” Neuendorf and Kumar (2015). Lacy and Riffe identify the measurement of inter-coder reliability as a strength of quantitative content analysis, arguing that, if content analysts do not measure inter-coder reliability, their data are no more reliable than the subjective impressions of a single reader Riffe et al. (1993). Berelson classifies the uses of content analysis in accordance to the question they seek to answer, the communication paradigm they are commonly applied to and their overall purpose Berelson (1952):

Purpose	Element	Question
Analyse the communication antecedents	Source Encoding process	who? why?
Analyse the communications characteristics	Channel Message Recipient	how? what? to whom?
Analyse the consequences of communication	Decoding process	with what effect?

Table 2.2: Barelson’s classification of the uses of Content Analysis

The automated analysis of Web text

The Web has become one of the most powerful tools invented by man mainly because of 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 of text analysis has been created: Web Analytics, *“the measurement, collection, analysis and reporting of web data for purposes of understanding and optimizing web usage”* Association (2007). This PhD project focuses on a subgroup of these methods, metrics and analysis techniques called content analysis. The main focus of content analysis is to assess texts through the systematic quantification of its resources to deliver replicable and valid inferences Duriau et al. (2007).

Clustering and sentiment analysis methodologies appear appropriate to evaluate the content and structure of PILs. Clustering is the process of grouping together individual objects or elements that are more similar to themselves than they are to the elements outside the set (Clements (1954). Sentiment analysis is the process of quantifying the emotion-related words present in a text, and 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 et al. (2004). Both of these methodologies have been extensively employed in the analysis of Web text, as it will be explored in the following sections.

Bidirectional encoder representations from transformers for language understanding

BERT is a language representation model that can be used for content analysis. It was developed by Google in 2018 and it is used to assign intent labels to the user queries. It uses a transformer-based machine learning technique for natural language processing and was pre-trained by Google on two tasks: language modelling and next sentence prediction.

It has been used in qualitative content analysis in psychological online counselling where a system of over 50 categories to analyse counseling conversations was developed requiring the manual labelling of 10,000 text passages Grandeit et al. (2020). It has also been used to model Italian social media language Polignano et al. (2019) where it was reported that training the BERT base took 11 days in the GPU while BERT large took 22 days. Thus, the resources needed to employ this model were deemed to be outside of the scope of this project.

Sentiment Analysis

Sentiment analysis can be defined as the process of categorizing the polarity of a text Qu et al. (2004)Qu et al. (2004),where polarity refers to the attitude of the text in the positive, negative and sometimes the neutral scale Stone et al. (1968). This process systematically identifies, extracts, quantifies and studies affective states and subjective information present in a text Volcani and Fogel (2006) and can be automated to evaluate attitudes such as:

- Judgements/evaluations: assess the overall perception of a product or 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 PhD research, sentiment analysis is employed to evaluate the polarity of PILs on the negative-positive scale and the proportion of emotional words on 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 on the patients' decision.

To do this, the NRC Emotion Lexicon version 0.92 Mohammad and Turney (2013) that contains the emotional relationships for 14,245 words has been employed. An emotion lexicon is used in sentiment analysis creating a list of words and their relationships with emotion categories. In the NRC Lexicon these relationships correspond to the central emotion categories on 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 (1872); Darwin et al. (1872) assumed that the evolutionary 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 fast reactions during emergency events Darwin (1998). These basic emotions would therefore be a basic component in the interpersonal interactions as they could signal imminent actions or intentions to others Darwin (1998) . 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 of differentiating between them. In addition, emotions nearer to each other on the wheel 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.

Cluster Analysis

As previously mentioned, clustering relates to the idea of grouping objects in accordance with a similarity metric. This methodology was first implemented not in Web analytics but in anthropology to evaluate the quantitative expression of cultural relationships Clements (1954) and has been adopted by practically all research disciplines in response to the need to process huge amounts of raw study 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 $\forall x, y \in C, \forall z \notin C \to d(x, z) < d(x, y)$, where $d : SxS \to R$ is a distance function between the elements in S.

The exact formation of the clusters clearly depends on the precise 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 being invented by researchers seeking to adapt the analysis method to the particulars of their data. These general approaches to defining a cluster can be categorized into Xiong et al. (2009):

- Evaluating connectivity distances

- Making a graph interpretation
- Employing statistical distributions
- Analysing density regions
- Structuring based on membership and attributes
- Employing mean vectors

In addition, these 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 to any cluster, to also belong to parent cluster and finally to have no restrictions on the number or type of clusters an element can belong to. In this PhD research, the words in the PILs are defined as the elements of the analysis while the degree of similarity is given by the number of words that two documents share. By seeking relationships between the appearance of certain words or combinations of words with readability metrics (see below) and measuring the understanding of facts about the trial the PIL describes, this project aims to provide trialists with valuable insights when writing their PILs.

2.4.4 Readability indexes

Readability can be defined as the ease with which a reader is able to understand a specific text fragment or document Mc Laughlin (1969). Readability analysis can generally be classified into two categories: content analysis, which focuses on the complexity of the vocabulary and syntax, and presentation analysis, which includes typographic elements like font size, line height, character spacing, and line length Harris and Hodges (1995). Interest in readability research can be tracked to Prof Sherman in the late 19th century who incorporated statistical analysis of Elizabethan and contemporary texts to compare their sentence structure Sherman (1893). His work demonstrated that statistical analysis could be employed in literary research, that shorter sentences and concrete terms help to understand written text, that speech is easier to understand than text and that overtime text, becomes easier to understand when it is more similar to speech:

> ”Literary English, in short, will follow the forms of standard spoken English from which it comes. No man should talk worse than he writes, no man should write better than he should talk.... The oral sentence is clearest because it is the product of millions of daily efforts to be clear and strong. It represents the work of the race for thousands of years in perfecting an effective instrument of communication.”–Prof Sherman, 1893

Several approaches have been used to assess the readability of textual documents, including:

- Text levelling, which is a subjective judgment commonly used to rank the reading ease of texts in areas where reading difficulties are easy to identify Fry (2002)
- Vocabulary frequency lists, which match word frequencies to readers skills Thorndike (1921), and
- Readability formulas, which try to assess multiple parameters to form a holistic measure of readability Gray and Leary (1935)

This PhD project focuses on the 5 most commonly employed readability measures for assessing documents intended to inform patients, which consider most of the metrics used to measure the readability of a document based on its textual characteristics, as shown in Table 2.3.

Aspects to determine document readability	Automated readability index	Flesch Kincaid Index	Gunning Fog Index	Coleman Liau index	Simple Measure of Gobbledygook
Total number of characters/letters	x			x	
Total number of words	x	x			
Average number of characters per word	x				
Average number of letters per 100 words				x	
Average number of sentences per 100 words				x	
Total number of sentences	x	x	x		x
Average number of words per sentence	x	x	x		
Total number of syllables		x			
Average number of syllables per Word		x			
Total number of difficult words					
Total number of complex words			x		
Ratio of sentences to complex words (Percentage)			x		
Number of words of 3 or more syllables			x		x
Average number of words of 3 or more syllables per 100 words			x		

Table 2.3: Readability indexes comparison table

The following sections discuss each of these readability indexes in more detail.

ARI

The Automated Readability Index (ARI) seeks to approximate the US grade (education level) needed to understand a text Smith and Kincaid (1970). It relies on two factors the ratio of characters per word and the ratio of words per sentence, using the following formula:

$$4.71(characters/words) + 0.5(words/sentences) - 21.43$$

where *characters* is the total number of letters in the document, *words* is the number of words obtained by counting the total number of spaces and *sentences* is the total number of sentences. The ARI index can classify English texts into 14 levels based on their reading ease, Table 2.4.

Score	Age	Grade Level
1	5-6	Kindergarten
2	6-7	First/Second Grade
3	7-9	Third Grade
4	9-10	Fourth Grade
5	10-11	Fifth Grade
6	11-12	Sixth Grade
7	12-13	Seventh Grade
8	13-14	Eighth Grade
9	14-15	Ninth Grade
10	15-16	Tenth Grade
11	16-17	Eleventh Grade
12	17-18	Twelfth grade
13	18-24	College student
14	24+	Professor

TABLE 2.4: Correspondence between ARI scores and school grade

The ARI readability index has been used to assess the barriers to health literacy Agness et al. (2008), the readability of paediatric patient information materials Swartz (2010), the fitness of purpose of patient information sheets Knapp et al. (2011a), and the identification of misleading strategies on online information Volkova and Jang (2018).

Flesch-Kincaid

Flesch-Kincaid refers to two readability tests that seek to assess how difficult an English language passage is to understand. The two tests are the Flesch Reading Ease and the Flesch-Kincaid Grade level, the tests employ the same factors to determine the readability of a text but incorporate different weights in their formulas Kincaid et al.

(1975):

$$FleschReadingEase = 206.835 - 1.015(tWords/tSentences) - 84.6(totalSyllables/tWords)$$

$$GradeLevel = .39(tWords/tSentences) + 11.8(totalSyllables/tWords) - 15.5$$

The results of both test are inversely correlated and were incorporated as part of US Navy research on technical manuals in 1978 Kincaid et al. (1981, 1988) and has since been widely adopted by researchers and policy makers in the USA McClure (1987). This readability index has been employed in the evaluation and development of PILs for clinical trials Sekhar et al. (2017); Suhaj et al. (2015); Roy et al. (2013); Kumaran et al. (2010) and the assessment of Web complaints on clinical procedures for low clarity data Rothrock et al. (2019); Terranova et al. (2012).

Gunning-Fog

The Gunning-Fog readability index was developed in 1952 by Robert Gunning to assess the readability of English texts Gunning et al. (1952) and focuses on the ratios between sentences and complex words to the total number of words. It considers a text with an index of 17 would require a reading skill of a collage graduate, while texts with an index below 8 should be comprehensible by most audiences Seely (2013). The Gunning-Fog formula is defined as:

$$Score = 0.4[(Words/Sentences) - 100(ComplexWords/Words)]$$

Table 2.5: Gunning-Fog Index

Score	**Corresponding US grade**
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eighth grade
7	Seventh grade
6	Sixth grade

Continued on next page

Table 2.5 – continued from previous page

Score	Corresponding US grade

The Gunning-Fog index has also been applied to the development and assessment of readability for PILs Suhaj et al. (2015); Rothrock et al. (2019); Munsour et al. (2017); O'Sullivan et al. (2020); Wong (1999); Liu et al. (2014) and online clinical information Elmadani (2019); Gill et al. (2012)

Coleman-Liau

The Coleman-Liau index seeks to simplify the process of automating the assessment of text readability. Contrary to other indexes, it focuses on ascertaining the boundaries between characters, words and sentences instead of syllables Coleman and Liau (1975). The principal advantage of this readability index is removing the need for optical character recognition or manual input and the higher accuracy and speed of computer programs in recognizing these elements Coleman and Liau (1975); Ley and Florio (1996). The Coleman-Liau formula is thus defined as:

$$CLI = 0.588L - 0.296S + 15.8$$

Where L is the ratio of letters per 100 words and S is the ratio of sentences per 100 words. The Coleman-Liau score approximates to the US grade level needed to understand the text.

This index has been used in the development and assessment of PILs readability Suhaj et al. (2015); Pringle et al. (2013); Knapp et al. (2011a), and web clinical information Martin and Pear (2015); Canfield (2020); Rothrock et al. (2019); Croft (2012); Dobbs et al. (2017); LeBrun et al. (2013); Murphy and Davis (1997).

Smog index

The Simple Measure of Gobbledygook (SMOG) readability index is a peer-validated instrument that seeks to estimate the necessary education level (based on the US grade system) needed to understand an English text. It is one of the most commonly used instruments employed to assess medical, clinical and health texts Ley and Florio (1996); Hedman (2008) intended for patients and general audiences, as the Smog grade has a 0.99 correlation with the grades of readers who had 100% comprehension of test materials with a standard error of 1.52 grades Mc Laughlin (1969). The Smog formula can be defined as:

$$Grade = 1.0430\sqrt{PollySilWords\ X\ (\frac{30}{Sentences})} + 3.1291$$

The SMOG readability index is commonly employed to assess and develop PILs for clinical trials Suhaj et al. (2015); Mumford (1997); Hamnes et al. (2016); MacDonald et al. (2010); Munsour et al. (2017), ascertain the readability of clinical policies and information intended for patients Robinson and McMenemy (2020); Sand-Jecklin (2007); Gill et al. (2012), and analyse the readability of online patient resources of information Dobbs et al. (2017); Elmadani (2019).

Chapter 3

General methodology

3.1 Overview

This chapter gives a general description of methodological approach of the primary studies that compose this thesis. This includes a brief description of the participants, recruitment, materials and experimental design. It also approaches the ethical considerations and data analysis needed for realizing this project. Further details specific to each individual study will be provided on the method section corresponding to its chapter.

3.2 Participants

3.2.1 Sample

Public participants were employed for three of the studies. In the second study, a stratified sample method was used to recruit 30 public participants were contacted via the Millennium Third Age Centre (3AC) a public charity based on Southampton. The third study recruited 30 public participants on the Amazon Mechanical Turk platform to review 4 current PILs for pragmatic trials on the UK. The third study also engaged 117 public participants on Amazon Mechanical Turk to reword sentences that were identified as being too hard to understand and an independent sample of 32 participants to find the best proposed revisions.

3.2.2 Sample Size

As there has not been a previous study on the necessary number of participants needed for a Public and Involvement group to assess a PIL because of the restrictions on researching PPI groups. The sample size of the experiments reflected initial conversations with PPI officers from the Faculty of Medicine and the Welcome Trust Clinic who advised that small groups (4-5 participants) are common when reviewing PILs for RCTs. Guidance for sample size varies with numbers ranging from 7 to 11 participants per focus group engaged on public involvement activities Francisca Caron-Flinterman et al. (2005).

3.2.3 Inclusion and exclusion criteria

The recruited participants were selected using a stratified sampling method to account for the differences in understanding and perception based on the education level of the participants. Three levels of education were pursued GCSEs, undergrad, and graduate. All participants must also be adults residing within the UK.

3.3 Recruitment

The recruitment to the primary studies was done through the crowdsourcing platform MTurk, the Millennium Third Age Centre, and the PPI officer from the University of Southampton Faculty of Medicine. A web platform was designed to collect, process and present public feedback, ”the web for public involvement”. The first study was an exploratory analysis on the associations between trial characteristics, recruitment and PIL readability and thus no participants were recruited. In the second study recruited public participants via the 3AC community and the PPI officer to read, comment and assess PILs. The third study published Human Intelligence Tasks (HITs) in the MTurk platform inviting public participants to revise and validate sentences that were deemed to be too hard to be understood by public audiences.

3.4 Methods

This research project employed an observational study to analyse the characteristics of PILs' text, and two non-clinical experimental studies to assess the effect of using crowdsourcing to revise and validate revision of PIL sentences that were found to be to hard to be understood by public audiences.

3.4.1 Observational studies

Clinical research studies are typically classified into two groups in accordance to their methodology into observational and experimental studies. Observational studies focus on observing the effect of risk factors, diagnostic tests, treatments or interventions without direct manipulation on whom is exposed to the effect MRC (2016) while experimental studies seek to understand the effect a treatment or intervention (independent variable) has in an outcome when that particular factor is manipulated often through the use of a control group.

The first study on this research project was an exploratory study of trial performance and PIL readability designed to identify inherent associations between PIL readability and trial characteristics. This study was based on a set of 58 PILs for trial supported by the HRA and its main objectives were to quantify the general readability of RCT PILs, investigate the use of words related to emotions in PIL text, and ascertain possible associations between readability and recruitment rates, use of emotive words and topics of research.

3.4.2 Experimental studies

3.4.2.1 Public involvement groups

As presented in the literature review chapter 2, Public involvement is currently defined by the NIHR-INVOLVE as "research that is carried out with or by members of the public rather than to, about or for them" for Health Research NHS (2018). This definition constrains the researcher in their approach as research cannot be carried out on those members of the public collaborating as part of a PPI group.

This research project sought to ascertain the effect of PIL readability on recruitment rates, the perception of quality and the understanding of the PILs' core aspects. This made necessary to incorporate process to measure the participants reading skill and understanding of PIL information and thus could not be considered PPI groups under the current definition. Two strategies were approached as potential solution for this dilemma, the employment of focus groups and the use of a crowdsourcing platform, which is explained in the following section. For the first strategy, the Third Millennium Age Centre (3AC) was approached to help recruit a stratified sample in accordance to their education level to read, comment, and assess PILs with low readability as focus groups of 5 persons. This would have been an identical approach to the expected for contributors in a PPI group.

In addition, each participant was asked to answer a multiple choice questionnaire after reading each PIL. The questionnaires were based on the topics identified as relevant by the EQIP guidelines Charvet-Berard et al. (2008)to ensure the quality of medical literature intended for patients. Finally, the participants were asked to complete a series of sentences based on the Cloze procedure to identify their reading skill level. This last two tasks were deemed to fall outside the intention of PPI as it is currently intended.

3.4.2.2 Crowd-sourcing

This project employed the Amazon Mechanical Turk (MTurk) platform to crowdsource the revision and validation of PIL sentences that were deemed to be too hard to be understood by public audiences. The MTurk platform enables the requester to publish Human Intelligence Tasks (HITs) to specific segments of the users based on demographic filters. This project associated MTurk HITs to the research Webtool to recruit participants for the following studies:

1. Collect public feedback on PILs with severe readability issues: The participants were asked to read and comment online 3 PILs with severe readability issues. In addition, they were asked to fill a demographic questionnaire, and to assess the information quality of the PILs based on the EQIP scale.
2. Revise PIL sentences that were deemed too hard to be understood by public audiences: The participants were asked to fill a demographic questionnaire, to complete three sentences in which every 5Th word had been replaced with a blank space and to reword 3 sentences that were deemed to be too hard to understand.
3. Validate proposed revisions to original PIL sentences: The participants were asked to fill a demographic questionnaire, and to grade 9 proposed revisions for each one of 27 PIL original sentences.

Each participant is assigned a worker id by MTurk, when a HIT is published the participant sees a description of the tasks, and a link to the research web page. Once the participant completes the task a code is given to them by the researcher to be submitted to MTurk. After the approval of the participant work the participant is rewarded the corresponding incentive for the task. This project used several subgroups (called batches) to recruit participants for each of the descriptive studies previously mentioned, as the commission to recruit more than 9 participants per batch increases the overall cost of the project.

3.5 Materials

3.5.1 Quality of information questionnaire

The Ensuring Quality of Information for Patients (EQIP) scale was published Beki Moult in 2004 has become a strong guidance for clinical researchers developing information for patients. Originally composed by 20 items which could assessed the compliance of the texts on reference details and quality criteria questions (as Yes, No and Partly), it has since been extended with 16 items to address document content, structure and the identification of essential information Charvet-Berard et al. (2008) to better assess the needs of medical research organizations. The development of the extended EQIP scale included rating the quality of 73 leaflets describing medical care procedures. It demonstrated very good inter-rater reliability (mean item-specific k statistic on 48 documents = 0.84) with an interclass correlation coefficient for the global score was 0.95, and a mean global conformity score on all items of 44 (range: 21–76, S.D. = 10). Given its proven value as an assessment tool for a large range of patient information resources, the extended EQIP scale has been adopted by several independent studies researching the quality of PILs intended for patients Charvet-Berard et al. (2008); Moult et al. (2004). This research project employed the extended EQIP guidelines to create questionnaires for assessing the understanding of essential trial information present on PILs for clinical trials. They were also used as a base to compare the results of content and thematic analysis on public feedback for PILs from clinical trials with poor recruitment. Further emphasis on the characteristics and impact of the scales are described in the corresponding sections of the literature review and the corresponding chapters of the primary studies they were used.

3.5.2 Web platform for public involvement in revising PILs

The first project objective is to create a Web-tool capable of collecting comments from public participants on PILs from low risk RCTs, which contain a similar coverage of topics as the expected comments from face-to-face PPI groups. To do this, it is essential to construct software modules to solve the following tasks:

1. Convert PDF to text
2. Clean text files
3. Convert text file to Json file (paragraph and sentence structures)
4. Upload to MySQL database
5. Create web page templates

 a Display study information

 b Demographic questionnaire

 c Create a participant id

 d Enable select and comment function

 e Present a EQIP questionnaire

 f Present a validation questionnaire

6. Upload the collected information to the database

The process of collecting information starts by receiving the PILs texts as PDF or Word documents that must be then converted into text files. The information in the text files must be cleaned from writing errors, water marks and other elements that may disrupt the analysis process. Afterwards, the text is parsed into paragraphs and sentences and uploaded to the database via a Json file. Once the information of the PIL is in the database, it can be accessed to conform web pages in which the public participants will join the study by reading the study information, answering a demographic questionnaire (which does not include identifiable information) and consenting to be part of the study. The study asks the public participants to read the text of the PIL and select parts that are not clear, after selecting some text a button appears to give the option to comment on the part. Once this task is concluded, the participants are presented with a multiple choice questionnaire about the PIL information based on the topics considered important to assess information quality by the EQIP guidelines. Finally, a short open question questionnaire on commonly used clinical terms is given to assess the participant familiarity with clinical research.

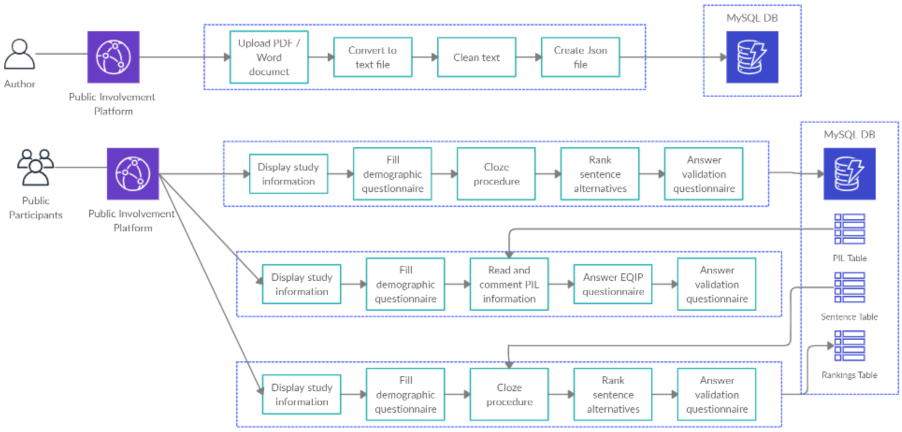

Author
Public Involvement Platform
Upload PDF / Word documet
Convert to text file
Clean text
Create Json file
MySQL DB
Public Participants
Public Involvement Platform
Display study information
Fill demographic questionnaire
Cloze procedure
Rank sentence alternatives
Answer validation questionnaire
MySQL DB
Display study information
Fill demographic questionnaire
Read and comment PIL information
Answer EQIP questionnaire
Answer validation questionnaire
PIL Table
Sentence Table
Display study information
Fill demographic questionnaire
Cloze procedure
Rank sentence alternatives
Answer validation questionnaire
Rankings Table

FIGURE 3.1: Web platform flow diagram.

3.6 Research design

How to design and deliver information to members of the public who are invited to participate in clinical trials and what information should trials communicate have been identified as top research question as current as 2018 Healy et al. (2018). This reflects the findings of multiple independent studies in the last two decades which have consistently found issues in RCT PILs Moult et al. (2004); Nicholls et al. (2009); Escudero-Carretero et al. (2013), in spite of the great effort made by UK health organizations in developing guidelines HRA (2014-03-03, 2017-01-17, 2020); MRC (2016); MHRA (2016) to ensure a high level of information quality is given to patients who are invited to participate in clinical trials. Thus, the main objective of this research project is to determine if a Web platform and text analysis techniques can become a novel solution to this issues by helping to identify, revise and validate PILs' sentences that are too difficult to understand by general audiences. This project was divided into three primary studies:

1. Identifying the characteristics of PILs' text that may be related to poor recruitment on RCTs.
2. Analysing the feedback given to PILs by public participants
3. Assessing the use of crowdsourcing, content analysis and readability metrics to identify, revise and validate readability issues on PILs by using a Webtool.

The following subsections provide a brief outlook on the methods used in each study. Further detail is provided on the methodology section of the corresponding study chapter.

3.6.1 PIL characteristics related to poor recruitment

The first step in this approach to help solve the issues previously mentioned was determined to be finding how the text of PILs with poor recruitment differs from the text commonly designed to be understood by general audiences. To do this, the project compared the sentence structure, word usage, topics, use of emotion and jargon presence in PILs when to texts designed to be understood and attract the attention of general audiences. . Figure 3.2 shows the flow diagram for selecting eligible PILs.

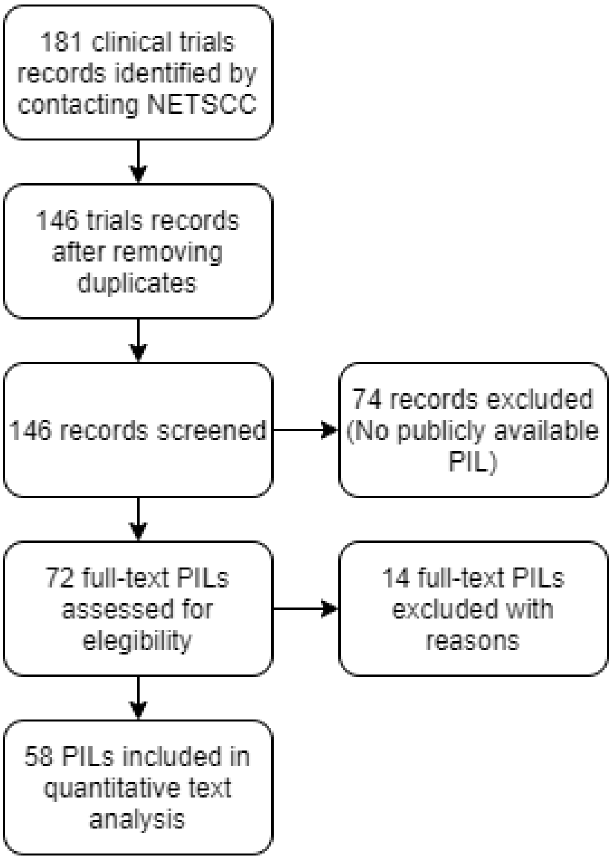

181 clinical trials records identified by contacting NETSCC
146 trials records after removing duplicates
146 records screened
74 records excluded (No publicly available PIL)
72 full-text PILs assessed for elegibility
14 full-text PILs excluded with reasons
58 PILs included in quantitative text analysis

Figure 3.2: Flow diagram for the selection of eligible PILs for text analysis.

3.6.2 Public feedback on PILs with severe readability issues

The second step in the research was to assess the feedback PPI groups gave to PILs that presented severe readability issues. This study analysed the themes and topics of comments given by public participants on PILs from RCTs founded by the NIHR HTA program between 2006 and 2009 that presented severe readability issues. The readability of each document was determined by the following factors:

1. The average readability score from 5 readability indexes Coleman-Liau, SMOG, ARI, Gunning-Fog and Flesch-Kincaid.
2. The PIL presented grammar, punctuation, or spelling mistakes
3. The presence of sentences with more than 15 words.
4. The presence of passive voice.
5. The presence of jargon words based on NHS guidelines

The underlying hypothesis of this study was that analysing the public feedback on PILs that presented severe readability issues would contain a good coverage of the common readability issues from the last primary study findings. The focus of this analysis is to determine the association between the participants' understanding of the PIL information, their perception of the quality of the information and the topics, and themes approached by their comments.

The participants were classified based on their education level into three groups, GCSE, undergrad and graduate to be able to assess the differences in performance and perception based on their education. The thematic analysis of the participant comments was used to validate the covering of the readability issues based on the EQIP guidelines, which is a validated instrument to ensure the quality of clinical information intended for patients or general audiences.

3.6.3 Employing crowdsourcing, content analysis and readability metrics to identify, revise and validate PIL readability issues

The final step on the research project is to analyse the use of crowdsourcing, content analysis and readability metrics to identify, revise and validate PIL readability issues. This include several descriptive sub-studies to:

1. Analyse the feedback of PIL authors
2. Identify and visualize readability issues via a Webtool
3. Assess the performance of MTurk participants who engage in revising PIL sentences that are too hard to be understood by public audiences
4. (a) Assess the effect of sentence difficulty on the time needed to revise the sentence
   (b) Assess the readability improvement of the proposed revisions
5. Assess the performance of MTurk participants who engage in validating revisions to PIL sentences that are too hard to be understood by public audiences.

To do this, a Web platform was created to analyse the text of PILs, identify sentences that require higher readability skills than the recommended for public audiences, capture public feedback from participants recruited from Amazon Mechanical Turk (MTurk), employ crowd-sourcing to revise and validate options to those sentences. The Web for Patient and Public Involvement on Revising RCT PILs (wppi.soton.ac.uk/) was thus developed to help realize the main objectives previously stated. In the first instance, it was necessary to ascertain if a Web platform could be used to collect public feedback with similar coverage as face to face PPI groups, which required the implementation of a system capable of capturing the PIL text, transforming it into a digital representation as a web page capable of storing the participant feedback on the PIL content, this process is detailed in Figure 3.3.

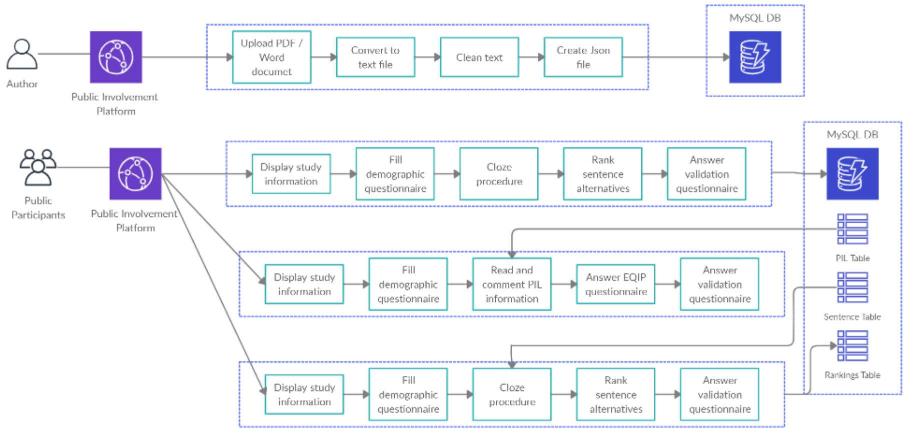

Author
Public Involvement Platform
Upload PDF / Word documet
Convert to text file
Clean text
Create Json file
MySQL DB
Public Participants
Public Involvement Platform
Display study information
Fill demographic questionnaire
Cloze procedure
Rank sentence alternatives
Answer validation questionnaire
MySQL DB
PIL Table
Display study information
Fill demographic questionnaire
Read and comment PIL information
Answer EQIP questionnaire
Answer validation questionnaire
Sentence Table
Rankings Table
Display study information
Fill demographic questionnaire
Cloze procedure
Rank sentence alternatives
Answer validation questionnaire

Figure 3.3: Flow diagram for the creation of web pages to capture public feedback on RCT PILs.

Chapter 4

Study I: Characteristics of PIL text

4.1 Overview

The first step of this research project was the analysis several PILs content to determine which characteristics could be linked to recruitment rates into RCTs. This analysis employed content and thematic analysis to determines the association between RCT recruitment rates and the sentences structure, level of emotive content, participant demographics and topic of research of the analysed PILs. This analysis sought to highlight the characteristics of PIL text that are associated to readability issues and poor recruitment, and it is used as a base for implementing a Web platform t capable of identifying, revising and validating readability issues on the third study (6).

Quantitative content analysis focus on frequency counts has made it an essential tool to explore the insights of news articles, where the size of the corpus makes infeasible to employ manual exploration and coding Kracauer (1952). Thus, many news agencies have adopted readability indexes to check the reading ease of their content based on their target audiences Sumpter (2001). News articles were gathered from the BBC, Daily Mail and Hello Magazine which target populations BBC-Bitesize (2018); Statista (2018); Magazine (2018) from different education levels, to anchor the results of the readability analysis.

4.2 Introduction

This study main hypothesis is that the current structure and content of RCT PILs is inhibiting their intended function to help inform participants who are invited to participate in clinical trials. The base of this hypothesis are the independent studies which have found serious readability issues on PILs, the current belief of main research stakeholders (nurses, clinical researchers) that PILs do not have an impact on the patient decision and the main approach to maintain an objective and formal tone when most psychology literature relates learning and memory to emotion.

4.3 Aims and objectives

The main aims of this study were:

1. Objectively quantify the general readability on RCT PILs
2. Objectively quantify the emotive content on PILs
3. Investigate the relationship between PIL readability and RCT recruitment rates
4. Investigate the relationship between PIL readability and the presence of emotive words
5. Investigate the relationship between PIL readability and the approached topics on clinical research

4.4 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.

4.4.1 Sample

58 Patient Information Leaflets (PILs) from RCTs supported by the NIHR program between 2000-2014 were collected and analysed.

Sample Size

The target for this study was to collect a diverse set of PIL documents for inviting participants into RCT trials. As presented in Chapter 2.4.4 the literature found the analysis of PILs to be commonly limited to single areas of research and with numbers of documents between 2-36. Agness et al. (2008); Swartz (2010); Knapp et al. (2011b); Karamitros et al. (2017). This study focused on PILs from trials supported by the NIHR as the response from commercially supported trials was lacking or the PILs were not available due to internal regulation. This limits the generalisability of the study results.

4.4.2 Ethics

This study focused on analysing the textual characteristics of publicly available PILs intended to inform prospective participants of essential details when invited to participate in clinical trials. No participant information was analysed, collected or stored. No need for ethical approval was deem necessary.

4.4.3 Procedure

Obtaining PIL text

Information on 181 trials that were supported by the NIHR program between 2000-2014 was provided by the NETSCC. The publicly available PILs from these trials were collected by following the procedure shown in Figure 4.1, where 74 records were excluded because they lacked a public available document and another 14 because they were not intended to inform the direct participant (i.e. PILs intended for children, non-consenting adults or cohorts). The PDF documents of the PILs were then processed into text documents employing an online OCR platform as detailed in Figure 4.2.

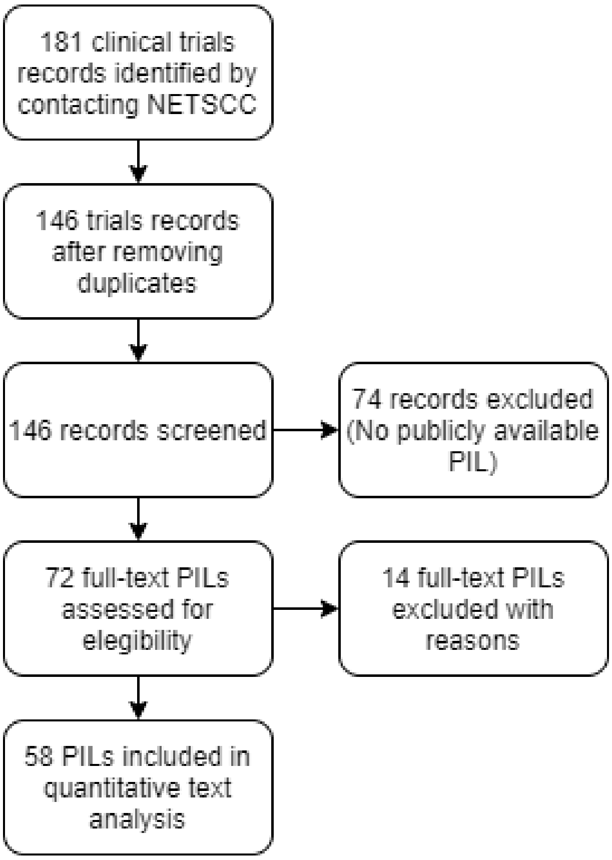

181 clinical trials records identified by contacting NETSCC
146 trials records after removing duplicates
146 records screened
74 records excluded (No publicly available PIL)
72 full-text PILs assessed for elegibility
14 full-text PILs excluded with reasons
58 PILs included in quantitative text analysis

FIGURE 4.1: Flow diagram for the selection of eligible PILs for text analysis.

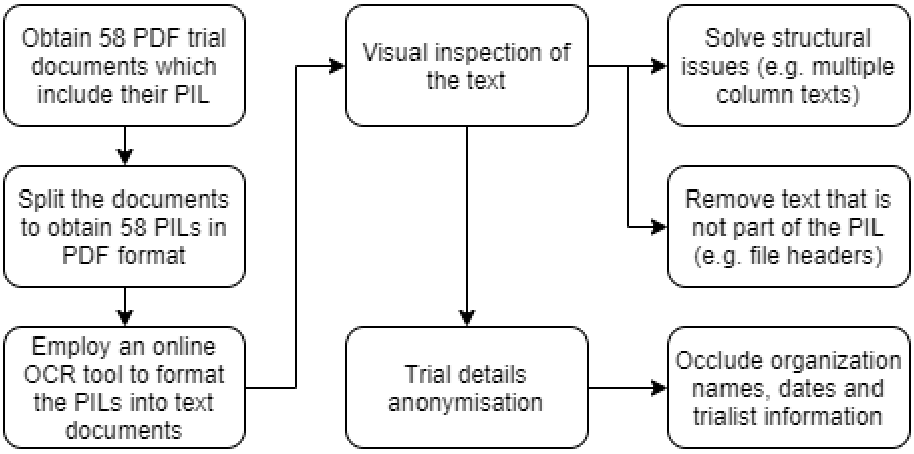

Obtain 58 PDF trial documents which include their PIL
Split the documents to obtain 58 PILs in PDF format
Employ an online OCR tool to format the PILs into text documents
Visual inspection of the text
Trial details anonymisation
Solve structural issues (e.g. multiple column texts)
Remove text that is not part of the PIL (e.g. file headers)
Occlude organization names, dates and trialist information

FIGURE 4.2: Flow diagram for extracting the PIL text from PDF trial documents.

Analysing PIL readability

Create a sentence vector representation Employ readability indexes to assess each PIL sentence Determine the overall readability of the document Identify sentences with poor readability scores

Analysing PIL emotive content

Create a word matrix representation for each PIL Employ NRC EmoLex to quantify the density of emotive words present in the text

Analysing PIL topics

create weighted lists of PILs words Employ cluster analysis to classify the words Discard clusters of words which appear on all documents Identify clusters which are correlate with recruitment rates Explore the themes and topics associated to these words

4.4.4 Statistical analysis

This was an exploratory study to identify inherent associations between recruitment to RCTs, the trials' characteristics, and the PIL readability. Multiple regression models were used including diverse characteristics. Table Figure 4.1and Figure 4.2 include a complete list of the explored characteristics.

4.5 Results

Multiple linear regression models were used to explore the associations between the trials capacity to recruit patients, the lexical characteristics of the PIL text, and the trial characteristics. Tables 4.1, 4.2, 4.4 and 4.3 introduce the variables explored.

Trial Settings
1. Sample size
(a) planned [scalar]
(b) Recruited [scalar]
(c) Amendments to sample size [nominal-dichotomous]
(d) Percentage recruited [scalar]
(e) Successfully recruited at least 80% [nominal-dichotomous]
2. Project start date
(a) planned [ordinal]
(b) Actual [ordinal]
3. Duration
(a) planned [scalar]
(b) Actual [scalar]
4. Type of care [nominal]
5. Type of setting [nominal]
6. Recruitment centres [nominal]
(a) planned [scalar]
(b) Actual [scalar]
(c) Number of recruitment centres who recruited patients [scalar]
7. ICD-10 disease classification [nominal]
8. HRCS codes [nominal]
9. Trial type [nominal]

TABLE 4.1: Independent trial variables analysed

One of the main objectives sought by this research was exploring the association between the trial characteristics and the readability of documents intended to inform patients who were invited to participate in those trials. The study results confirm the existence of association between PIL readability and trial factors, which are presented in Section 4.5.2 were consistent with the study hypothesis that the current proportionate approach to seeking consent requiring less oversight to documents from low risk trials could fail in identifying even severe readability issues in those PILs.

PIL characteristics

1. PIL lexical characteristics
   (a) Number of words [scalar]
   (b) Characters per word [scalar]
   (c) Syllables per word[scalar]
   (d) Words per sentence [scalar]
   (e) Complex words (%) [scalar]
   (f) Words associated to an emotion or sentiment (%) [scalar]

Table 4.2: PIL lexical variables analysed

The composition of the PILs was a factor considered by this research. The exploration of the variables detailed in Table 4.2 is presented in Section 4.5.1 where it was observed a large variance in the composition of the PILs lexical structure. This study found no direct association between the lexical aspects of the documents and PIL readability or RCT recruitment, but noted those variables are inherent parts for the readability indexes described in Table 4.3. These readability indexes gave consistent scores when assessing the PILs and were found to be associated to RCT recruitment as shown in Section 4.5.2.

PIL Readability

1. PIL readability scores
   (a) SMOG score [scalar]
   (b) Flesch Kincaid score [scalar]
   (c) ARI score [scalar]
   (d) Coleman Liau score [scalar]

Table 4.3: Independent readability variables

This study explored the association between the PILs emotive content, the documents readability and the recruitment to trials in Section 4.5.2, the explored variables are described in Table 4.3. It was found that there is a significant association between the presence of emotive words in PILs, recruitment to RCTs, and readability scores.

PIL Emotive Content
1. Positive words (%) [scalar]
2. Negative words (%) [scalar]
3. Words related to anger (%) [scalar]
4. Words related to anticipation (%) [scalar]
5. Words related to disgust (%) [scalar]
6. Words related to fear (%) [scalar]
7. Words related to joy (%) [scalar]
8. Words related to sadness (%) [scalar]
9. Words related to surprise (%) [scalar]
10. Emotive content (%) [scalar]

Table 4.4: Independent emotive variables

The following sections summarize the inherent associations identified by this exploratory study in greater detail.

4.5.1 Trial types

The analysis of 58 PILs was carried to analysed the associations between trial settings, PIL characteristics and PIL readability described in the previous section. Table 4.5 gives a summary of the number of documents classified in each area based on the trial type, classification, research area and disease code. It was found some categories lacked enough representation to meet the requirement for significance in the regression models of Section 4.5.2 and were merged with other subcategories to be included in the analysis as presented in Yates method Moore (1996).

Table 4.5: Trial characteristics.

RCT characteristics	**n=58**
Trial type	
Surgery	9
Drug	15
Diagnostic	4
Service delivery	1
Psychological therapy	3
Devices	8
Physical therapy	18
UKCRC code	
Evaluation treatment	36
Disease management	3
Health services	1
Other	18
Research setting	
Hospital	31
Community	2
GP	9
Multi-setting	11
ICD-10 disease code	
Respiratory	8
Muskolo-skeletal	6
Ganito-urinary	5
Health status	3
Mental	9
Nervous	3
Circulatory	6

Categories with less than 5 representatives were merged together where possible, in case no other category with less than 5 elements was present they were excluded from this set of analyses. The joined categories were: Diagnostic-Service delivery-Psychological therapy, Disease management-Health Services and Health status-Nervous. The community setting category with only 2 elements was excluded from these analyses.

PIL readability

Large differences were observed in the lexical structure of each individual PIL as shown in Tables 4.6 and 4.7 when the PILs were classified by type and area respectively.

Table 4.6: PIL characteristics based on RCT type (Intervention Category)

Type	**N**	**Words**		**Characters per Word**		**Syllables per Word**		**Words per Sentence**		**Complex Words (%)**	
		Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev
Physical therapy	18	1840	1385	4.63	0.14	1.53	0.06	20.18	3.24	12.7	2.26
Drug	15	1670	809	4.65	0.2	1.53	0.08	21.02	3.81	12.98	2.69
Surgery	9	1830	1519	4.63	0.19	1.54	0.09	21.73	3.87	13.2	3.4
Diagnostic	8	902	405	4.63	0.14	1.52	0.05	20.32	2.62	13.34	2.18
Devices	8	1197	554	4.66	0.13	1.52	0.04	22.06	5.23	12.32	1.66
Total	58	1576	1110	4.63	0.16	1.53	0.07	20.92	3.68	12.89	2.44

A one way ANOVA test was used to assess the lexical variables within the RCT type and research area groups with the exception of the number of words, which in all cases the observed variances of the documents shown in Tables 4.6 and 4.7 were deemed to be above the needed level to ensure the significance of the results. The results shown no statistically significant difference in the number of characters per word, syllables per word, words per sentence or the percentage of complex words for trial types(p=0.991, p=0.97, p=0.721, p=0.92) or research areas (p=0.3, p=0.114, p=0.946, p=0.117). The large variance in the number of words is consistent with the proportionate approach to seeking consent in which documents for low risk trials are designed to be significantly shorter.

Table 4.7: PIL characteristics based on their RCT research area (ICD-10)

Category (ICD-10)	**N**	**Words**		**Characters per Word**		**Syllables per Word**		**Words per Sentence**		**Complex Words (%)**	
		Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev
Mental	9	1695	1377	4.6	0.08	1.49	0.03	21.6	4.03	11.84	1.52
Nervous	6	1148	502	4.54	0.09	1.5	0.04	20.36	3.95	12.38	1.66
Circulatory	6	1897	1871	4.6	0.06	1.5	0.02	20.77	5.92	12.38	1.35
Respiratory	8	2120	1731	4.72	0.19	1.56	0.08	22.14	3.23	13.84	2.85

Continued on next page

Table 4.7 – continued from previous page

ICD-10 category	N	Words		Characters per Word		Syllables per Word		Words per Sentence		Complex Words (%)	
		Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev	Avg Num	Std Dev
Muskolo-skeletal	6	1012	426	4.72	0.11	1.57	0.06	20.15	2.34	14.66	2.46
Ganito-urinary	5	1411	619	4.69	0.27	1.58	0.11	20.82	4.73	14.56	4.04
Other	18	1545	625	4.62	0.18	1.52	0.07	20.54	3.19	12.28	2.25
Total	58	1576	1110	4.63	0.16	1.53	0.07	20.92	3.68	12.89	2.44

Also a one sided ANOVA test was used to analyse differences in the scores reported by the Smog, Flesch-Kincaid, Gunning-Fog, Coleman-Liau and ARI readability indexes. All readability indexes produced consistent scores as observed in Tables 4.8 and 4.9, no statistically significant difference was observed by the test between the mean scores reported in the groups for trial type (p=0.904, p=0.801, p=0.865, p=0.96, p=0.722 respectively) or research category (p=0.374, p=0.766, p=0.62, p=0.253, p=0.785 respectively).

Table 4.8: PIL readability scores based on their RCT type

Type	SMOG index		Flesch-Kincaid index		Gunning-Fog index		Coleman-Liau index		ARI index	
	Mean Score	Std Dev	Mean Score	Std Dev	Mean Score	Std Dev	Mean Score	Std Dev	Mean Score	Std Dev
Physical therapy	11.71	0.89	10.31	1.33	12.76	1.39	10.46	0.85	10.46	1.61
Drug	12.00	1.45	10.71	1.92	13.17	2.21	10.44	1.21	10.97	2.32
Surgery	12.16	1.44	11.03	1.90	13.51	1.92	10.45	1.13	11.22	2.04
Diagnostic	12.32	0.67	10.71	0.41	13.60	0.65	10.63	1.07	10.77	0.55
Devices	11.97	1.23	10.90	2.11	13.35	2.37	10.75	0.72	11.53	2.81

The results shown in Tables 4.8 and 4.9 also indicate that most PIL documents need above the maximum recommended of 10^{th} grade education when creating understandable documents for general audiences and that this is a common issue in trials across diverse settings. The readability of the documents was found to be significantly associated to the type of trial as described in Section 4.5.2.

Table 4.9: PIL readability scores based on their RCT research area (ICD-10)

Category (ICD-10)	**SMOG index**		**Flesch-Kincaid index**		**Gunning-Fog index**		**Coleman-Liau index**		**ARI index**	
	Mean Score	Std Dev	Mean Score	Std Dev	Mean Score	Std Dev	Mean Score	Std Dev	Mean Score	Std Dev
Mental	11.73	1.16	10.47	1.75	12.98	2.00	10.18	0.70	11.03	2.09
Nervous	12.12	0.16	11.03	1.09	13.60	0.69	10.09	0.72	11.36	1.70
Circulatory	11.68	0.96	10.24	2.11	12.93	2.26	10.45	0.39	10.61	2.86
Respiratory	12.55	1.42	11.35	1.87	14.05	2.01	11.01	1.15	11.88	2.11
Muskolo-skeletal	12.35	0.63	10.77	0.81	13.53	0.85	10.99	0.67	10.86	0.98
Ganito-urinary	12.47	2.05	11.20	2.78	13.68	2.97	10.85	1.59	11.10	2.98
Health status	11.12	0.47	9.22	0.89	11.87	0.83	9.84	0.29	8.94	1.13
Other	11.63	0.96	10.41	1.33	12.67	1.43	10.35	1.01	10.60	1.76

PIL recruitment by research area and trial type

The recruitment to RCTs was measured by the percentage of the initial sample that was successfully recruited. To maintain high significance trialist seek to maintain at least 80% recruitment. This study focused on assessing the association between the type of trial and area of research with the readability of the PILs. The study results shown that being a trial on evaluating the effects of a drug was significantly associated to the recruited sample as it is described in Section 4.5.2.

Table 4.10: Average percentage recruited (planned sample) by RCT type

Type	N	Mean	Std Dev
Physical therapy	18	84.61	43.84
Drug	15	62.45	31.85
Surgery	9	54.82	36.90
Diagnostic	4	74.64	23.97
Service delivery	1	107.38	
Psychological therapy	3	92.74	34.31

The observed data presented in Tables 4.10 and 4.11 shows that trials from drug and surgery studies have lower recruitment rates, and that when reclassifying by disease studied trials on respiratory, muscolo-skeletal and ganito-urinary studies have better recruitment rates.

Table 4.11: Average percentage recruited (planned sample) by RCT research area (ICD-10)

Type	N	Mean	Std Dev
Mental	9	65.43	32.61
Nervous	3	65.04	60.13
Circulatory	6	68.20	51.54
Respiratory	8	79.61	29.67
Muskolo-skeletal	6	88.99	48.98
Ganito-urinary	5	74.64	32.37
Health status	3	54.65	48.13
Other	18	70.66	33.05

PIL emotive content

This study results shown that the emotional content of the documents was significantly associated to the PIL readability as described in Section 4.5.2. Tables 4.12, 4.13 and 4.14 describe the mean and variance of the percentage of words related to each emotive category present in the documents when classifying by trial type.

Table 4.12: Average percentage of emotive content present in PILs by their trial classification

Type	**N**	**Positive**		**Negative**		**Anger**		**Anticipation**	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Physical therapy	18	4.70	1.67	1.39	0.83	0.40	0.40	1.48	0.58
Drug	15	6.00	2.00	1.76	1.09	0.38	0.21	1.87	0.81
Surgery	9	3.26	1.34	1.46	0.58	0.37	0.30	1.09	0.37
Diagnostic	4	3.82	2.46	1.22	1.18	0.28	0.34	1.44	0.82
Service delivery	1	2.72		0.16		0.05		0.98	
Psychological therapy	3	5.64	2.56	1.09	0.28	0.14	0.06	1.13	0.64
Devices	8	5.03	1.10	1.54	0.36	0.49	0.30	1.56	0.43

Table 4.12 shows PILs contain comparatively large percentage of words associated to positive sentiments. This result was consistent with a previous MSc research study Santos (2017) that found PILs contained more words related to positive sentiments than articles from Hello Magazine, Daily Mail and BBC.

Table 4.13: Average percentage of emotive content present in PILs by their trial classification Part 2

Type	**N**	**Disgust**		**Fear**		**Joy**		**Sadness**	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Physical therapy	18	0.40	0.49	1.07	0.63	0.65	0.60	1.00	0.59
Drug	15	0.62	0.48	1.23	0.73	1.13	0.75	1.05	0.59
Surgery	9	0.42	0.35	1.57	0.64	0.45	0.53	0.97	0.56
Diagnostic	4	0.29	0.29	1.37	1.36	0.45	0.22	0.70	0.63
Service delivery	1	0.00		1.52		0.49		1.25	
Psychological therapy	3	0.05	0.04	0.45	0.32	0.50	0.36	0.73	0.43
Devices	8	0.41	0.27	1.01	0.40	0.61	0.30	1.03	0.29

Excluding words related to positive sentiments it was observed that the PILs contained a low percentage content related to emotive words and that there were large variances between the documents. The percentage of words related to joy and negative emotions were found to be significant factors when predicting the recruited proportion of the RCT samples as detailed in Section 4.5.2.

Table 4.14: Average percentage of emotive content present in PILs by their trial classification Part 3

Type	**N**	**Surprise**		**Trust**	
		Mean	Std Dev	Mean	Std Dev
Physical therapy	18	0.41	0.20	1.92	0.72
Drug	15	0.46	0.18	2.23	0.82
Surgery	9	0.27	0.12	1.71	0.48
Diagnostic	4	0.29	0.16	1.93	1.22
Service delivery	1	0.16		3.21	
Psychological therapy	3	0.32	0.23	2.49	1.19
Devices	8	0.50	0.29	2.27	0.63

4.5.2 Associations

Readability aspects associated to recruitment

The recruited proportion of the initial sample of the RCTs was shown to be significantly associated to PIL readability, the emotion content of the documents and the trial settings. In particular, a decrease in the Flesch readability scores and percentage of negative words shown to correlate with increments in the recruited proportion. On the other hand, not being a drug intervention, having more words related to joy and having resized the sample increased the expected recruitment. When limiting the factors to individual readability scores, it was shown that the Coleman-Liau readability scores significantly correlated to the recruited samples. A similar analysis of the percentage of emotive words found no significant associations when employing individual linear regression models.

A multiple linear regression model was used to predict the recruited proportion of the planned sample for the RCTs. R^2 for the overall model was 88.1%, with an adjusted R^2 of 82.6%, $F(5,11) = 16.229,\ p < 0.001$ .The model formula was defined as:

$RecruitedProportion = 231.34 - 2.909 FleschScore + 0.547 JoyWords + 42.58 SampleIncr - 28.706 DrugIntervention - 0.46 NegWords$

Association between readability and trial settings

When analysing the factors associated to the readability scores, the study found the readability scores were significantly associated to recruitment and trial settings. Specifically, increments in the Smog readability scores positively correlated with the proportion of sample recruited in the trials and being a diagnostic intervention, while having at least 10 recruiting centres and not being a trial to evaluate a treatment were associated to lower readability scores. A Pearson correlation was used to identify characteristics associated to the readability of the PILs. The multiple linear regression model identified significant associations between the SMOG readability scores and the recruited proportion of the planned sample, having less than 10 centres, being an evaluation of a treatment and having a diagnostic intervention. The R^2 for the overall model was 87.6% with an adjusted R^2 of 82%, $F(5,11) = 15.568\ p < .001$. The model formula was defined as:

$SmogScr = 11.873 + 0.013 RecProp - 0.929 NumCentres10 - 0.988 EvalTreat + 1.148 InterDiag$

Association between readability and recruitment

In addition, a linear multivariate regression model was employed to explore the inherent associations of the PILs lexical characteristics on recruitment to RCTs. These characteristics shown in Table 4.15 are commonly used to determine how difficult general audiences find to understand a document.

Table 4.15: Readability characteristics.

Readability characteristics
Number of words
Number of characters
Number of syllables
Number of sentences
Number of complex words
SMOG index
Flesch-Kincaid index
Gunning-Fog index
Coleman-Liay index
ARI index

Linear regression models were run to predict the percentage of the recruited sample from the number of words, characters, syllables, sentences, and complex words present in their PILs. These models also included the readability scores for each PIL by five indexes SMOG, Flesch-Kincaid, Gunning-Fog,Coleman-Liau, and ARI. A regression model found a Coleman-Liau scores statistically significantly predicted the percentage recruited of the planned sample, $F(1, 56) = 5.25,\ R^2 = .086,\ p = .026,\ Percentage\ recruited = -46.992 + 11.33(Coleman - Liau\ score)$.

Finally, the percentage of emotive content present in the leaflets was not found to be directly a significant factor to predict the percentage of the recruited sample when modeled by themselves. But they were found to be significant when added to a multivariate regression model with the readability scores of the PILs as previously stated in this section 4.5.2.

Linear regression models were employed to explore the inherent associations of each emotive category to recruitment into RCT trials. Table 4.16 list the individual variables explored.

Table 4.16: Independent variables.

Emotive content
Percentage of words associated to positive sentiments
Percentage of words associated to negative sentiments
Percentage of words associated to anger
Percentage of words associated to anticipation
Percentage of words associated to disgust
Percentage of words associated to fear
Percentage of words associated to joy
Percentage of words associated to sadness
Percentage of words associated to surprise
Percentage of words associated to trust

The linear regression models to predict the percentage of the recruited sample from the independent variables showed none of the individual variables added statistically significantly to the prediction at $p < 0.001$.

4.6 Summary

The literature has consistently found readability issues on PILs intended to inform participants who are being recruited to RCTs during the last two decades. Solving these issues has become a priority area of research for UK organizations linked to health research. The main objective of this study was to assess the readability and characterize PILs supported by the HRA with publicly available documents. The focus of this study was to identify associations between the characteristics of trials and PIL documents.

Our analysis showed significant correlation between the readability score (Smog index) and the recruited proportion of the planned sample ($p = 0.015$), having 10 or less centres recruit participants ($p < .001$), and being a trial for evaluating a treatment($p < .001$). We also found significant associations between the recruited proportion of the planned sample and the Flesch-Kincaid reading score ($p = .006$), the amount of words related to joy ($p = .002$) and negative ($p < .001$) emotions, having the sample size increased ($p = .001$) and having a drug intervention ($p < .001$). When grouped by type or research area the PILs demonstrated large variances in their number of words, characters, syllable, sentences and complex words, but all the readability indexes gave consistent measurements on the education level needed to understand the PILs (in all cases requiring more than 10th grade).

A comparison of the emotion present in the PILs with news articles from the BBC, Daily Mail and Hello Magazine to anchor the scale with a gradient of texts intended to inform different segments of public audiences. Here the results shown that the PILs contained significantly more words related to positive emotions and significantly less related to negative emotions than any of the other groups. On the other hand, the PILs included significantly less words related to other emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) assessed.

These results shown that the readability of PILs and their proportion of emotive content can be associated to the recruitment rates on their trials. The heavily skew percentage of positive words and lack of words related to other emotions could be a result of the current guidance on maintaining a respectful tone of voice while engaging the patient, but literature on emotion and behaviour indicate readers in emotive states may find difficult to pay attention, understand and remember information which lacks emotive stimulus or which greatly diverges from their emotive state.

Chapter 5

Study II: Public Feedback to PILs

5.1 Introduction

This primary study focus on analysing the public feedback of a public involvement group on RCT PILs. Clinical researchers are encouraged to engage Patient & Public Involvement (PPI) groups as part of the proportional HRA model to seeking consent. Given the high cost of engaging PPI groups when following the guidelines for paying PPI fees and the additional work it conveys, many clinical researchers of pragmatic trials choose to not engage them. Thus, pragmatic trials PILs mostly designed by following PIL templates, which focus on covering the legal requirements but not on ensuring the understanding of the essential trial information.

Public Involvement in research is defined as research carried out 'with' or 'by' members of the public rather than 'to', 'about' or 'for' them (NIHR-INVOLVE, 2018). This compels an interesting dilemma: when PPI groups are employed to assess the quality of PILs intended to inform potential participants of an RCT, they can be employed as reviewers, give comments, quality assessments and suggestions for revisions, but the researcher can not directly investigate their effect or assess their understanding of PIL the information as that would make them participants in research instead of research contributors. Knowing if the information is being understood as intended by the researcher is one of the core aspects needed to produce high-quality PILs. Thus, this study analyses the association between the feedback commonly given by public participants and their understanding of the information present in the PILs.

The main hypothesis of this study is that these limitations to engaging PPI groups are part of the cause of the severe quality issues on PILs that have been reported in the last two decades Healy et al. (2018); Gillies et al. (2014); Reinert et al. (2014); Knapp et al. (2011a). Additionally, an hypothesis was formed that such issues in the quality of the information partly account for the poor recruitment rates of up to two thirds of the publicly funded studies Raftery et al. (2015).

5.2 Aims and objectives

The main aims of this study were:

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

5.3 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. The general methodology was described in 3, the following sections focus on the specific details for this study.

5.3.1 Sample

30 public participants were recruited through the Millennium Third Age Centre (3AC) with a stratified sampling method based on their education level (GCSEs, Undergrad, and Graduate). The 3AC a public charity organization based on Southampton. Given the current restriction on researching PPI contributors there is no current framework on sample size for PPI groups, the sample size on this study was determined by following the guidance for PPI groups (5 persons per group) given by the PPI officer from the University of Southampton Faculty of Medicine.

The participants were asked to read comment and assess a subset of 4 PILs selected from the previous study. This PILs were selected on the basis of being the most difficult to read, thus maximising the participants' opportunity to comment. It was noted that this PILs also correspond to RCTs with publicly available PILs supported by the NIHR HTA program between 2006-2009 and recruitment rates below 15

5.3.2 Ethics

As this study sought to determine associations between commonly given feedback to PILs and the level of understanding provided by the PILs, the participants would be deemed subjects of research and the recruitment could not be provided by the university PPI officer.

5.3.3 Procedure

Before the session

An add was posted with at the 3AC centre inviting public participants to the study. Each potential participant would be given a leaflet with the information about the study by the 3AC staff containing a summary about the study tasks and objectives:

1. Read and comment on 4 RCT PIL
2. Assess the PIL information quality
3. Answer a multiple choice questionnaire to determine their understanding of the information.

Potential participants were then provided with a consent form to join the study, and asked to fill a demographic questionnaire. A selection of the interested participants was made based on their education level: GCSEs or below, undergraduate, and graduate or above. The selected participants were provided with the details of the study session.

All participant information was anonymized by assigning each participant with a randomly generated alphanumeric id. The multiple choice questionnaire was developed based on the extended EQIP guidelines on developing high-quality information for clinical information intended for patients (Charvet-Berard, Chopard, & Perneger, 2008). The extended EQIP scale is an instrument that has demonstrated very good inter-rater reliability on assessing the quality of information intended for patients and general audiences.

5.3.3.1 The experiment

This experiment consisted on a series of experiments as detailed in Figure 5.1, where participants were presented with 4 PILs from RCTs supported by the NIHR HTA program between 2006-2009. The PILs' text will be provided to the participants as a printout document 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

The participants were grouped in teams of 5 people with similar education levels, who engaged each other while reading, and commenting the PILs information, but each participant was also asked to individually answer a multiple choice questionnaire about the PIL information and assess the information present in each PIL.

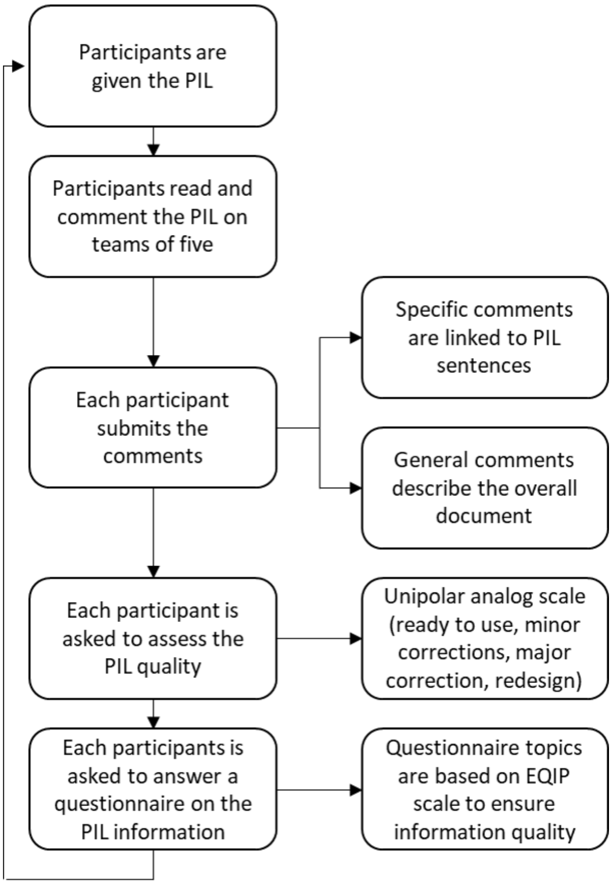

Participants are given the PIL
Participants read and comment the PIL on teams of five
Specific comments are linked to PIL sentences
Each participant submits the comments
General comments describe the overall document
Each participant is asked to assess the PIL quality
Unipolar analog scale (ready to use, minor corrections, major correction, redesign)
Each participants is asked to answer a questionnaire on the PIL information
Questionnaire topics are based on EQIP scale to ensure information quality

Figure 5.1: Flow diagram for the study tasks.

Participants were given 1 PIL and asked to read and comment it on their group. Each participant would then record their comments on the provided form. Two types of comments were recorded by the participants: comments to specific sentences or sections of the PILs and comments on the overall quality of the document. The qualitative assessment of the PIL information quality was recorded by employing an unipolar analog scale (no quality issues, minor issues, moderate issues, severe quality issues, should not be used). A multiple choice questionnaire based on topics identified as essential to creating high-quality information by the extended EQIP guidelines was used to assess the participants understanding of the PIL essential information after delivering the feedback commonly expected of a PPI group.

5.3.4 After the experiment

After the study session the participants were asked to fill an event participation form for the 3AC charity, and were rewarded £10 for their PPI contribution.

5.4 Statistical analysis

5.4.1 PIL understanding

The understanding of the core aspects of the trials was assessed by employing a multiple choice questionnaire, these questionnaires were based on the topics identified by the extended EQIP guidelines as essential in determining the quality of information intended for potential patients. The EQIP scale is an instrument that has demonstrated high inter-rater reliability for assessing information quality on clinical information intended for patients or general audiences.

A cumulative odds ordinal logistic regression model was used to assess the association between understanding and the participants' perception of PIL quality. The model is a multivariate extension to the logistic regression model that applies to dichotomous dependent variables McCullagh (1980). It is mainly used on when a priory interest is the comparison of the individual categories of the response variable Ananth and Kleinbaum (1997).

The ordinal logistic regression model has been found to be appropriate for use in comparative studies of quality of life where the scales can have ordered categories grouped on the basis of a continuous variable or discrete categories with an underlying order Abreu et al. (2008). This study seeks to determine the odds of participants being more likely of failing to identify (minor or no issues) readability issues in the documents based on their scores (percentage of correct answers obtained).

5.4.2 Participant feedback

Thematic analysis was employed to analyse the comments collected from the participants. These comments could be linked to specific sections or sentences (specific comments) or discuss general issues or characteristics of the document (general comments).

A sample of 20% of the comments was randomly selected to verify the inter-rater reliability, which was independently coded by a second researcher for each of coding phases. Both raters showed at least moderate agreement for all the coding phases. Figure 5.1 and Figure 5.2 show a summary of the reliability scores under the NCBI-NIH rating model for reliability McHugh (2012).

Coding	Percentage Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha (nominal)
NVivo	72.40%	0.572	0.586	0.579
Sentiment	69%	0.459	0.463	0.468
Topic	75.90%	0.592	0.594	0.599
Purpose	75.90%	0.536	0.546	0.544

TABLE 5.1: Inter-rater reliability scores

Coding	N Agreements	N Disagreements	N Cases
NVivo	21	8	29
Sentiment	20	9	29
Topic	22	7	29
Purpose	22	7	29

TABLE 5.2: Inter-rater reliability scores

In addition, an unipolar analog scale was used to measure the qualitatively assess the participants perception of overall quality of the documents, Figure 5.2.

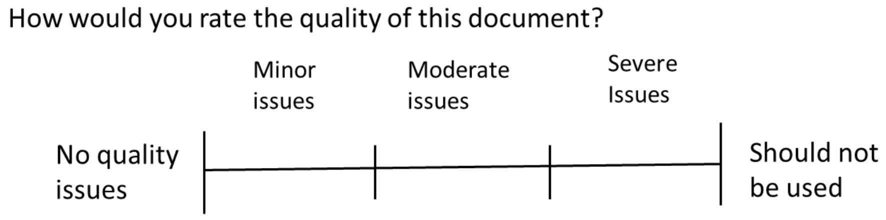


FIGURE 5.2: Scale to assess quality perception.

5.5 Results

5.5.1 Participant characteristics

Table 5.3 gives a summary of the participant characteristics. Six of the recruited participants did not complete the minimum required study tasks and where excluded from the study.

A stratification sampling method based on their education level was used to subdivide the participants into three groups: GCSEs, Undergrad and Graduate education.

Participants characteristics	n=24
Age (years)	
18-30	7
31-44	8
45+	7
Not stated	2
Sex	
Male	7
Female	14
Not stated	3
Education level	
GCSEs or below	9
Undergrad	7
Graduate or above	8
Origin	
UK	13
European	4
International	7

Table 5.3: Participants characteristics

5.5.2 Qualitative analysis

A 4-cycle thematic analysis Saldaña (2015) was used on the participants' comments to assess their type, purpose, emotional spectrum and topic.

NVivo coding

The process started with an NVivo coding cycle to obtain general categories of for the participants' comments, four categories referring to the purpose in which the comments were made were identified in this cycle and two referring to their type:

- Type of comments:
  - General comments were comments given by the participants that were not associated to any particular section of the document
  - Specific comments were comments associated to sections, phrases or specific wording of the document.
- Purpose of the comments:
  - 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/Disapproval ("It was good")

In the first instance the 240 comments were divided into general comments (nm=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 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.

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:

Emotion coding

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".

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 tables Table 5,Table 6 summarize the emotive classes for comments given by the reviewers.

Code	Description	Examples
"I would like more information"	This type of comment is a request for further explanation on specific terms, concepts, procedures or other information present in the document	"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"
"Instead of using ..."	These are indications to change the format, structure, or design of the reviewed document, procedure or information	"-However- better word than -but-" "Simplify -relative merits- to advantages" "Can we not see it on a website? Or email for it?"
"Too hard to understand"	Mostly composed by general comments that give qualitative observations on the document information, composition and readability.	"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"
"It was good"	This are general comments that endorse the quality of the document.	"I like the friendly writing style" "This was a good one"

Table 5.4: NVivo codes

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").

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. Seven 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"

Code	Description	Examples
Positive	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	"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."
Neutral	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. Sixty-eight comments from a set of 159 specific comments were deemed as 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)"
Negative	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.	"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"

Table 5.5: Sentiment codes (specific comments)

Motif coding

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

Code	Description	Examples
Positive	Forty-nine general comments from a set of 81 were deemed as positive. These comments tended to express qualitative assessments of the documents.	"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."
Neutral	Only Six general comments were deemed to be neutral. They provided specific advice even when not associated with a particular section of the document.	"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"
Negative	Twenty six 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.	"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."

Table 5.6: Sentiment codes (general comments)

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).

In this cycle, motif coding was used 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. The first set of comments in this cycle are related to how the purpose of the studies are explained to potential participants in the trials. Common highlights from the reviewers included 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 group 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:

> 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

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.

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 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. All the topics approached by the EQIP scale were represented in the comments given by the participants with the exclusion of those referring to the presentation of visual elements and the consent form as the participants were only given the PILs’ text.

In addition, participants' comments included aspects not directly assessed by the scale which demonstrated the study generated a good coverage of the relevant readability issues in PILs.

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. These 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”

5.5.3 Quantitative analysis

The purpose of the quantitative analysis was to assess the participants’ perception of PIL quality and their understanding of the PIL information. This analysis found no significant association between the participants reported PIL quality and their understanding of the information, but it found a clear distinction in the perception of the quality based on the participants' level of education with higher rates of lower educated participants correctly identifying the PILs contained severe and moderate readability issues.

Participant understanding

The understanding of the PIL information was assessed by the average score obtained by the participants on a multiple choice questionnaire on the PIL content. The questionnaire were developed based on the proposed topics by the extended EQIP scale. The answer sheets to the first PIL questionnaire of participants who fail to answer at least 80% of the questions were excluded from the analysis of that PIL results. 5.7 shows the mean scores of the participants based on their education levels.

Group	Correct Ans PIL1 (%) (SD)	Correct Ans PIL2 (%) (SD)	Correct Ans PIL3 (%) (SD)	Correct Ans PIL4 (%) (SD)
GCSEs	16.8 (7.1)	22.1 (8.7)	23.5 (10.4)	27 (8.9)
Undergrad	63.7 (12)	52.9 (12.7)	48.2 (12.8)	50.6 (19.8)
Graduate	72.3 (12.1)	62.2 (12.2)	58 (9.8)	65.5 (18.1)

Table 5.7: Avg percentage of correct answers per reviewer after reading and commenting each leaflet

Perception of information quality

A unipolar analog scale was used to observe the participants' perception of information quality (No quality issues, minor issues, moderate issues, severe issues, should not be used). Table 5.8 show the distribution of assessments given by the participants to each PIL based on the their education levels. The participant assessment of the quality of the document was found to vary in accordance with their education level, with lower educated participants reporting higher severity of readability issues.

Group	Do not use (%)	Severe quality issues (%)	Moderate quality issues (%)	Minor quality issues (%)	No quality issues (%)
PIL 1					
GCSEs	14.3	14.3	42.9	14.3	14.3
Undergrad	0	16.7	16.7	66.7	0
Graduate	0	14.3	14.3	57.1	14.3
PIL 2					
GCSEs	0	0	83.3	0	16.7
Undergrad	0	0	16.7	66.7	16.7
Graduate	0	28.6	0	42.9	28.6
PIL 3					
GCSEs	0	0	83.3	0	16.7
Undergrad	0	20	0	60	20
Graduate	0	14.3	28.6	28.6	28.6
PIL 4					
GCSEs	0	40	60	0	0
Undergrad	0	20	0	40	40
Graduate	0	14.3	14.3	42.9	28.6

Table 5.8: Distribution of participants (%) who assessed quality

Association between understanding of PIL information and quality perception

A cumulative odds ordinal logistic regression with proportional odds model was used to assess the correlation between the percentage of correct answers obtained by the participants and their perception of the PIL quality.

GCSE Group

The assumption of proportional odds was not met, as assessed by a full likelihood ratio test comparing the fit of the proportional odds location model to a model with varying location parameters, $\chi2(2) = 8.316, p = .016$.

The Pearson goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi2(20) = 15.935, p = .721$. The deviance goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi2(20) = 14.910, p = .782$. The final model did not statistically significantly predicted the dependent variable over and above the intercept-only model, $\chi2(1) = .712, p < .399$.

No statistically significant association was found between the participant understanding of the PIL and their perception of the PIL quality, $Exp(B)_{[odds\ ratio]} = -.042$, 95% $CI[-.141, .057]$, $\chi^2(1) = .686, p = .408$

Undergrad Group

The assumption of proportional odds was not met, as assessed by a full likelihood ratio test comparing the fit of the proportional odds location model to a model with varying location parameters, $\chi 2(2) = 6.689, p = .035$.

The Pearson goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi 2(20) = 21.198, p = .386$. The deviance goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi 2(20) = 19.982, p = .459$.The final model did not statistically significantly predicted the dependent variable over and above the intercept-only model, $\chi 2(1) = .108, p < .742$.

No statistically significant association was found between the participant understanding of the PIL and their perception of the PIL quality, $Exp(B)_{[odds\ ratio]} = .990$, 95% $CI[.936, 1.048]$, $Wald\ \chi^2(1) = .118, p = .731$

Graduate Group

The assumption of proportional odds was not met, as assessed by a full likelihood ratio test comparing the fit of the proportional odds location model to a model with varying location parameters, $\chi 2(2) = 29.772, p < .05$.

The Pearson goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi 2(23) = 29.090, p = .177$. The deviance goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi 2(23) = 34.785, p = .055$.The final model did not statistically significantly predicted the dependent variable over and above the intercept-only model, $\chi 2(1) = .020, p < .889$.

No statistically significant association was found between the participant understanding of the PIL and their perception of the PIL quality, $Exp(B)_{[odds\ ratio]} = .996$, 95% $CI[.947, 1.047]$, $Wald\ \chi^2(1) = .025, p = .875$

Association between understanding of PIL information and participant feedback

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.

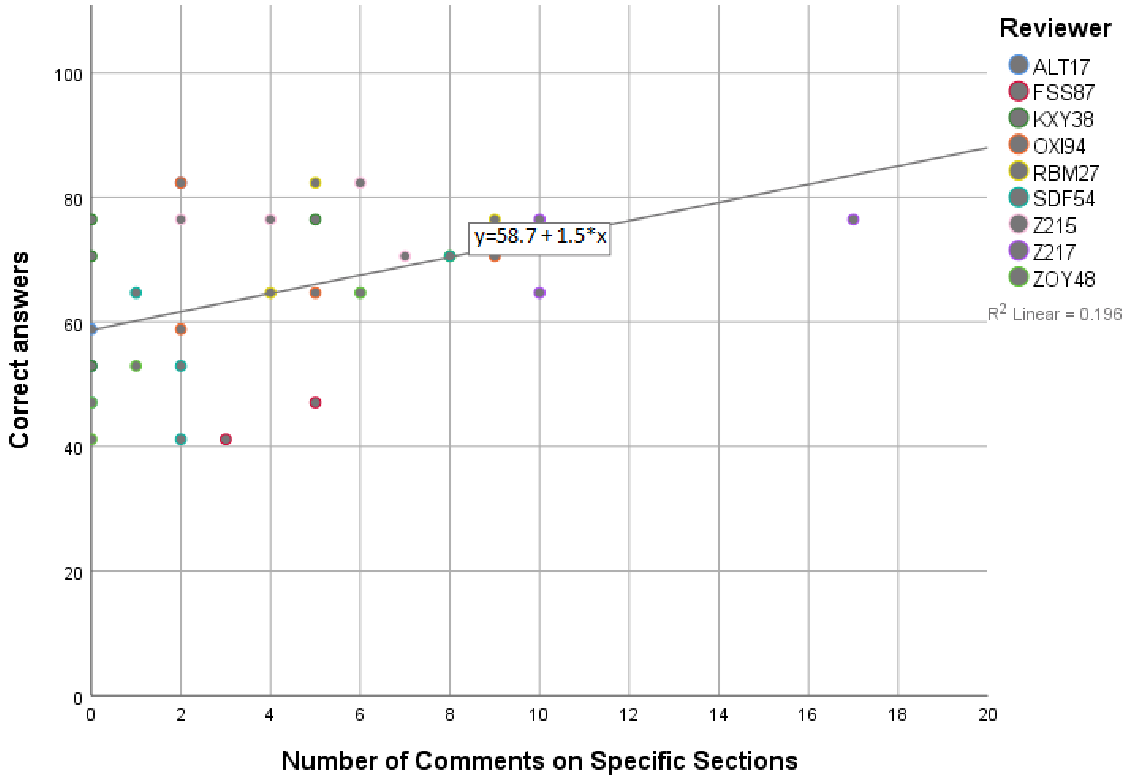


FIGURE 5.3: Association between the number of specific comments and the percentage of correct answers given by the reviewer.

Figure 5.3 shows a positive association between the number of comments associated to specific sections of the PIL given by a participant and the number of correct answers they obtained when assessed on the PIL information. This means the specific comments can be used to indirectly observe the participants' level of understanding of the PIL information, i.e. to be able to give a specific comment the participants need to have a certain level of understanding of the information or producing specific comments induces a learning process in the participant increasing their understanding of the information. This can also support the observed decrease in the dispersion of the data points as the number of comments increases.

The linear regression formula in Figure 5.3 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).

5.6 Summary

The main focus of this study was to explore the inherent associations between participant understanding and their perception of information quality in public audiences. The literature in the area has consistently found severe issues with the readability of PILs for recruiting participants into RCTs, and that most RCTs struggle to successfully recruit at least 80% of their planned samples. Even though this has been a priority area of research for UK health research organizations in the last two decades, this study confirmed the issues expressed by the literature on current PIL leaflets and explored novel approach to understanding these issues.

We studied 3 groups of public participants based on their education level (GCSEs, Undergrad, Grad) who assessed 4 PILs with significant readability issues based on four readability indexes (ARI, Gunning-Fog, Smog, and Flesch-Kincaid). No significant association was found between their qualitative perception of the quality of the PILs information and their understanding of the information for any of the groups. The groups presented distinct impressions of the PILs quality in all cases, but the Undergrad and Graduate groups had significantly more participants who only recognized minor o no quality issues with the leaflets ($x_1 = 15.5\%, x_2 = 77.5\%, x_3 = 67.9\%$)). This means a simple assessment of the PIL where a group of recruited participants use a scale to determine if it can be used is an unreliable method to validate the quality of RCT PILs.

A thematic analysis of the participant feedback, shown two distinct types of comments: qualitative comments on the overall document (general comments) and comments focused on specific sections of the PIL (specific comments). The amount of specific comments given by a participant was found to be associated to their understanding of the information (p<0.01). The intention of the comments also greatly varied between general and specific comments, where general comments were based subjective assessments which confronted or praised the authors; the specific comments tended to be neutral toned requests for more information, clarification or modification of the PIL text. In addition, the topics approached by the participants were found to closely follow the expanded EQIP scale for assessing the quality of medical information intended for patients.

Our thematic analysis shows that employing participants to review the PILs intended to recruit participants into RCTs can help identify issues in the documents, but the process must be guided to focus the participants on giving specific feedback instead of emotive context. Furthermore, the analysis shown some areas that are not currently not directly assessed by the EQIP scale like the clarity of the randomization process, the inclusion of repeated or redundant information, and the presence of bad writing (grammar, spelling, and punctuation mistakes or the use of inappropriate terms).

These results have brought forward insights on considering some aspects of the current HRA model of “proportionate approach to seeking consent”, which seeks to separate clinical research and public involvement, where public involvement is defined as "research being carried with or by members of the public rather than to, about or for them". Given that the current guideline to identify and address information quality issues on PIL is to engage a Patient and Public Involvement (PPI) group, this definition inhibits the researcher to assess the participants' understanding of the information, having to relay directly on the participant perception of quality. As this study shows, these assessments may not reflect the PIL capacity to inform the participants.

Chapter 6

Study III: Employing crowdsourcing, content analysis and readability metrics to identify, revise and validate PIL readability issues

6.1 Introduction

The previous studies have corroborated the presence of severe issues in the information presented to potential participants of pragmatic trials presented by independent authors and health institutions on the last two decades (Healy, et al., 2018),(Gillies, Huang, Skea, Brehaut, & Cotton, 2014), (Reinert, et al., 2014), (Knapp, Raynor, Silcock, & Parkinson, 2011), (Knapp, Raynor, Silcock, & Parkinson, 2011),(Charvet-Berard, Chopard, & Perneger, 2008). Even if the current HRA proportionate approach to seeking consent tries to solve these issues by inviting clinical researchers on engaging PPI groups and creating guidelines and frameworks to develop leaflets intended to inform potential RCT participants, this issues remain today as top research priority questions (Healy, et al., 2018). This research has shown that the current approach to PPI do not facilitate the clinical researchers from pragmatic trials to engage with PPI groups, where the recommended fees are above £25 per person and there is no requirement by the clinical researcher to do so.

It is the main hypothesis of this study that employing these techniques will result in an improvement of the readability of PIL and that creating a Web platform to identify, reword and validate PIL sentences that are too hard to understand that does not requires large investments in time or finances from the clinical researcher could greatly improve the information quality of PILs for pragmatic trials.

6.2 Aims and objectives

This primary study has as main aim to assess several approaches on increasing PIL readability by:

1. Employing readability indexes to identify text that is too hard to be understood by public audiences
2. Employing a Webtool to visualize the public feedback associated to each section of the PIL
3. Employing MTurk to crowdsource the revision of PIL sentences
4. Employing MTurk to crowdsource the validation of proposed revisions of PIL sentences.

6.3 Studies

This study was composed of three main parts: 1) the assessment of the readability of 3 PIL currently in use for pragmatic trials, 2) The assessment of employing crowdsourcing to revise PIL sentences that are too hard to be understood by public audiences and 3) The assessment of employing crowdsourcing to identify the best proposed revisions for each sentence. The following sections include the detail of the results of each part.

Five readability indexes were used to identify PIL sentences that were too hard to be understood by public audiences (required more than 9Th grade US education), Coleman-Liau, Gunning-Fog, SMOG, ARI, and Flesch-Kincaid. A sentence was deemed to be too hard if it exceeded 15 words in length and at least three of the indexes reported a score greater than 9Th grade. An average of 34% (s.d.=2%) of the PIL sentences were deemed to be too hard to be understood by public audiences. In all cases the overall readability of the PILs was greater than 10. There were no significant differences in the readability ratings given to each PIL by the indexes.

A group of 117 public participants crowd-sourced via the MTurk platform provided 677 revisions for a subset of 27 original PIL sentences. This subset was selected based on the sentence overall readability score (three levels) and their topic. The readability improvement on the proposed revisions when compared to the original sentences was found to be significantly associated ($R^2 = 0.407,\ p < .001$) to the amount of complex words and characters present in the sentence and the time the reviser expended doing the revision. The amount of time the participants expended on each revision was found to include extreme outliers were participants abandoned the task and concluded at later dates. This cases needed to be excluded.

In all the groups the time expended in revising the sentence was found to increase with the sentence difficulty. A multivariate linear model ($R^2 = 0.221,\ p < .001$) associated the total duration of the revision with the sentence difficulty, the number of characters in the sentence and its ARI and Gunning-Fog readability scores.

A one-way ANOVA test was used to identify statistically significant differences in the grades given by MTurk participants to a set of proposed revisions for original PIL sentences. 9 proposed revisions were randomly chosen for each sentence in a set of 27 original PIL sentences. The original sentences were classified into three levels of difficulty by their readability scores. The study collected 2,394 valid grade submissions for individual proposed revisions, from 32 participants. Only in two cases ($F(26, 505) = 2.76\ and\ F(26, 505) = 2.63\ both\ with\ p < 0.001$) the participants showed a significant preference for one of the proposed revisions, in both cases corresponded to a sentence with the highest difficulty level from different PILs.

6.3.1 Employing a Web platform to identify text that is too hard to be understood by public audiences and collect public feedback on PIL text.

A personalized website was created were public participants were presented the text of 3 PILs This study is a randomised experiment with 3 non-medical interventions with independent group.

Sample

30 public revisers 3 PIL authors

Sample Size

The target for this sub-study was to obtain a significant coverage of the issues present in the PILs, our previous study (presented in the previous chapter Section 5.5.2) indicated that 30 reviewers could provide an adequate coverage of topics for the issues present in the PILs.

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.

30 public participants were recruited via the crowdsourcing platform MTurk to assess the quality of the PIL information based on an EQIP questionnaire. 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 focused 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.

Results

PIL characteristics

In the first part, 30 participants were recruited by the MTurk platform to assess the quality of 3 PILs currently in use. The assessment was carried out using the extended EQIP questionnaire, a peer validated tool to ensure the quality of medical information, the Cloze procedure on a random selection of PIL sentences, and quantitative analysis of the PIL text via readability indexes. Tables 6.1 and 6.2 summarize the characteristics of the PILs researched in this study.

Table 6.1: PIL textual characteristics.

PIL characteristics	
PIL 1	
Sentences	133
Difficult sentences	46
Words	2594
Complex words	382
Characters	12405
Syllables	4102
PIL 2	
Sentences	60
Difficult sentences	19
Words	1186
Complex words	169
Characters	5560
Syllables	1836
PIL 3	
Sentences	136
Difficult sentences	47
Words	2601
Complex words	410
Characters	12654
Syllables	4204

Table 6.1 shows the text composition of the PILs contained some variance in its structure. Thus following analysis were necessary to determine the acceptability of its comparison. The first was an exploration of the overall expected reading skill needed to easily understand the text presented in Table 6.2.

Table 6.2: Mean PIL readability scores.

PIL readability	
PIL 1	
Readability score (ARI)	11^{th} Grade
Readability score (Coleman-Liau)	11^{th} Grade
Readability score (SMOG)	12^{th} Grade

Continued on next page

Table 6.2 – continued from previous page

PIL readability	
Readability score (Gunning-Fog)	13^{th} Grade
Readability score (Flesch-Kincaid)	11^{th} Grade
PIL 2	
Readability score (ARI)	11^{th} Grade
Readability score (Coleman-Liau)	11^{th} Grade
Readability score (SMOG)	12^{th} Grade
Readability score (Gunning-Fog)	13^{th} Grade
Readability score (Flesch-Kincaid)	10^{th} Grade
PIL 3	
Readability score (ARI)	11^{th} Grade
Readability score (Coleman-Liau)	12^{th} Grade
Readability score (SMOG)	12^{th} Grade
Readability score (Gunning-Fog)	13^{th} Grade
Readability score (Flesch-Kincaid)	11^{th} Grade

Table 6.2 shows that all indexes gave consistent scores for the PILs within 2 grades for the expected education level needed to easily understand the documents. The Gunning-Fog index was observed to give higher scores in all cases which may indicated a need to calibrate the consensus algorithm to automate the selection of particular sentences that are too hard to understand.

Automated identification of PIL sentences that are too hard to understand by general audiences

The scores of five readability indexes (ARI, Coleman-Liau, SMOG, Flesch-Kincaid and Gunning-Fog) were used to identify sentences that were too hard to be understood by general audiences. The procedure assessed a sentence to be too hard to be understood if three or more indexes scored the sentence above 9th grade. An average of 34% (sd=2%) of the PILs' sentences were deemed to be too hard to understand. The overall readability scores for each PIL are presented in Fig 6.1.

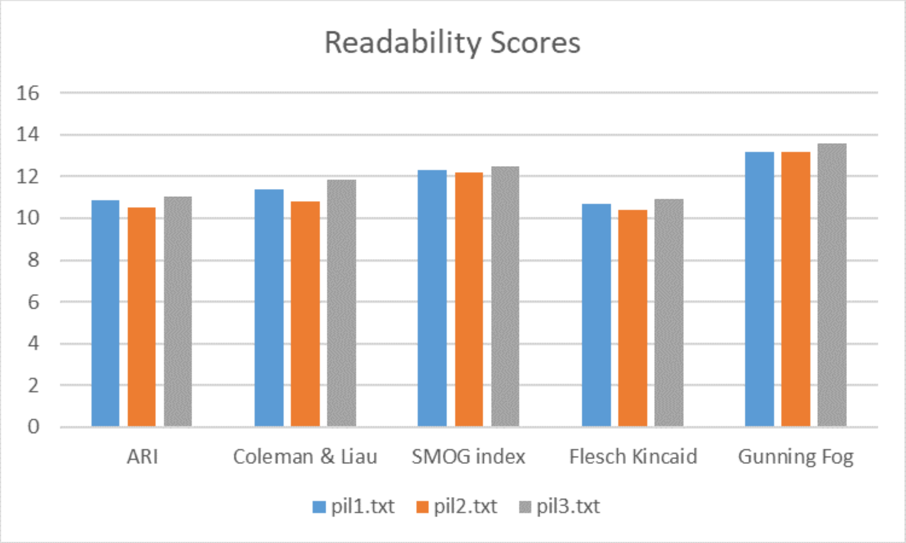


FIGURE 6.1: Readability scores per document.

Figure 6.1 shows that all the selected readability indexes gave similar scores to the three selected PILs. Which would mean all the indexes qualified the documents as being similar in their difficulty to be understood. Based on the previous results it was concluded that a comparison between sentences taken from these documents would be valid.

Issue coverage

Three aspects of the participants feedback were measured summarized in Table 6.3: the expected level of understanding of the document by general audiences based on the Cloze procedure score; the coverage of essential aspects to ensure information quality with the EQIP score; and the participant perception of document quality with a qualitative scale (Document ready to be used, needs minor corrections, needs major corrections or needs redesign).

The EQIP scale is a peer validated scale to ensure quality of information on clinical information intended for patients. Eqip recommendations are: For scores above 75% document is ready for use, above 50% can be used but you should consider doing a revision within 1-2 years, bellow 50% the PIL must be revised asap.

The Cloze procedure generates scores which are strongly correlated to the readers perception of the degree of difficulty in the narrative and the understandability of the document. It is suggested a minimum score of 57% is needed to ensure understandability.

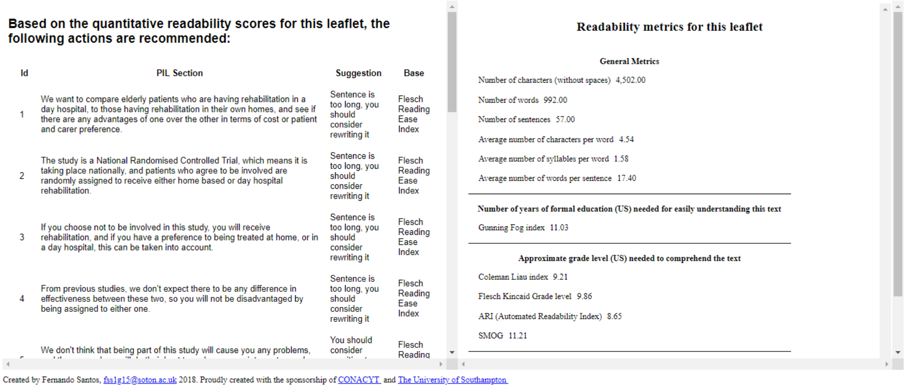
Table 6.3: Participants' feedback on the selected PILs.

Participants' feedback	
PIL 1	
Number of reviewers	27
Graded ready to be used	15%
Graded minor corrections	54%
Graded redesign is needed	23%
EQIP avg score	59%
Cloze procedure score	33%
PIL 2	
Number of reviewers	30
Graded ready to be used	30%
Graded minor corrections	60%
Graded redesign is needed	10%
EQIP avg score	74%
Cloze procedure score	53%
PIL 3	
Number of reviewers	26
Graded ready to be used	38%
Graded minor corrections	62%
Graded redesign is needed	0%
EQIP avg score	80%
Cloze procedure score	54%

The observed results in this section show that only one of the PILs could be considered ready to be use in its current version by EQIP and none cover the minimum score to ensure understandability base on the Cloze procedure. It was also noted that only in the first case were this scores greatly bellow the recommendation while the other two cases were near the borderline. In all cases the majority of the participants indicated that the current documents had only minor or no issues (69% PIL1, 90% PIL2, 100% PIL3). These results show the PILs analysed in this study contained less issues than those analysed in the previous chapter.

Feedback visualization

The designed Webtool has the capacity to identify specific sentences in the document in an automated manner and generate a report of the PIL readability for the PIL authors as shown in Figure 6.2.



Based on the quantitative readability scores for this leaflet, the following actions are recommended:

Id	PIL Section	Suggestion	Base
1	We want to compare elderly patients who are having rehabilitation in a day hospital, to those having rehabilitation in their own homes, and see if there are any advantages of one over the other in terms of cost or patient and carer preference.	Sentence is too long, you should consider rewriting it	Flesch Reading Ease Index
2	The study is a National Randomised Controlled Trial, which means it is taking place nationally, and patients who agree to be involved are randomly assigned to receive either home based or day hospital rehabilitation.	Sentence is too long, you should consider rewriting it	Flesch Reading Ease Index
3	If you choose not to be involved in this study, you will receive rehabilitation, and if you have a preference to being treated at home, or in a day hospital, this can be taken into account.	Sentence is too long, you should consider rewriting it	Flesch Reading Ease Index
4	From previous studies, we don't expect there to be any difference in effectiveness between these two, so you will not be disadvantaged by being assigned to either one.	Sentence is too long, you should consider rewriting it	Flesch Reading Ease Index
	We don't think that being part of this study will cause you any problems,	You should consider	Flesch Reading

Readability metrics for this leaflet

General Metrics

Number of characters (without spaces) 4,502.00

Number of words 992.00

Number of sentences 57.00

Average number of characters per word 4.54

Average number of syllables per word 1.58

Average number of words per sentence 17.40

Number of years of formal education (US) needed for easily understanding this text

Gunning Fog index 11.03

Approximate grade level (US) needed to comprehend the text

Coleman Liau index 9.21

Flesch Kincaid Grade level 9.86

ARI (Automated Readability Index) 8.65

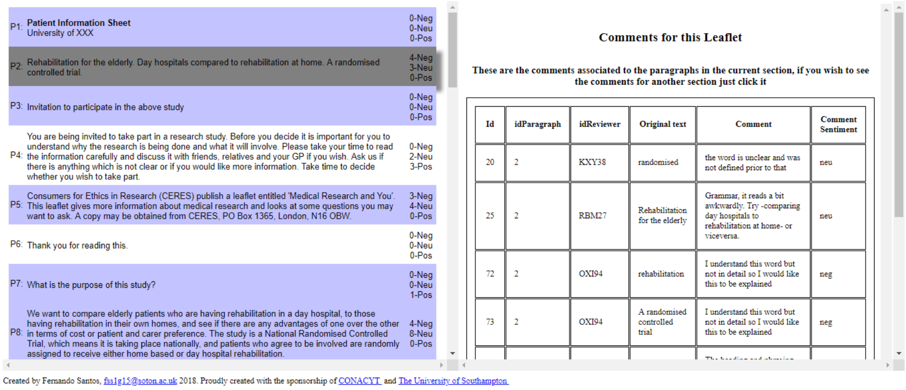
SMOG 11.21

Created by Fernando Santos, fss1g15@soton.ac.uk 2018. Proudly created with the sponsorship of CONACYT and The University of Southampton

Figure 6.2: Readability report.

This report would in a first instance give a provenance of why each specific sentence was chosen and link the revisions with known issues to help in further instances when revising other PILs. In addition, it would provide a base of the understandability of the current document that would help quantify the overall effect of the revision process in latter stages.

Figure 6.3 shows the visualization of the comments given by the participants to each section of the PIL. This would help the researchers identify specific sections were the participants had most issues and easily peruse all the comments given to that section.



P1:	**Patient Information Sheet** University of XXX	0-Neg 0-Neu 0-Pos
P2:	Rehabilitation for the elderly. Day hospitals compared to rehabilitation at home. A randomised controlled trial.	4-Neg 3-Neu 0-Pos
P3:	Invitation to participate in the above study	0-Neg 0-Neu 0-Pos
P4:	You are being invited to take part in a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take your time to read the information carefully and discuss it with friends, relatives and your GP if you wish. Ask us if there is anything which is not clear or if you would like more information. Take time to decide whether you wish to take part.	0-Neg 2-Neu 3-Pos
P5:	Consumers for Ethics in Research (CERES) publish a leaflet entitled 'Medical Research and You'. This leaflet gives more information about medical research and looks at some questions you may want to ask. A copy may be obtained from CERES, PO Box 1365, London, N16 OBW.	3-Neg 4-Neu 0-Pos
P6:	Thank you for reading this.	0-Neg 0-Neu 0-Pos
P7:	What is the purpose of this study?	0-Neg 0-Neu 1-Pos
P8:	We want to compare elderly patients who are having rehabilitation in a day hospital, to those having rehabilitation in their own homes, and see if there are any advantages of one over the other in terms of cost or patient and carer preference. The study is a National Randomised Controlled Trial, which means it is taking place nationally, and patients who agree to be involved are randomly assigned to receive either home based or day hospital rehabilitation.	4-Neg 8-Neu 0-Pos

Comments for this Leaflet

These are the comments associated to the paragraphs in the current section, if you wish to see the comments for another section just click it

Id	idParagraph	idReviewer	Original text	Comment	Comment Sentiment
20	2	KXY38	randomised	the word is unclear and was not defined prior to that	neu
25	2	RBM27	Rehabilitation for the elderly	Grammar, it reads a bit awkwardly. Try -comparing day hospitals to rehabilitation at home- or viceversa.	neu
72	2	OXI94	rehabilitation	I understand this word but not in detail so I would like this to be explained	neg
73	2	OXI94	A randomised controlled trial	I understand this word but not in detail so I would like this to be explained	neg

Created by Fernando Santos, fss1g15@soton.ac.uk 2018. Proudly created with the sponsorship of CONACYT and The University of Southampton

Figure 6.3: Comment visualization.

Finally a series of semi-structured interviews with 3 of the PIL authors researched. This was done to present the public feedback results on their results via the Webtool and receive feedback on the value it presents to them when designing PILs. All participants concluded that the Webtool presented value to PIL authors. They considered the visualization of public comments related to each section of a PIL to offer the most value and that public and patient feedback is of upmost importance to create high quality information. Also, everyone considered that keeping jargon understandable and controlling the length of the document are the two greatest challenges when developing a PIL. Other observations included the perception that current PILs have issues which must be resolved and that this issues currently affect the patients' decision to join RCTs. All authors expressed that just identifying and visualizing sentences, which are too hard to understand is not enough to help PIL authors, but that an idea on how to correct the sentences must be given.

6.3.2 Employing a web platform and crowdsourcing to revise PIL sentences that are too hard to be understood by public audiences

The results from this study demonstrated that crowdsourcing is a valid option to revise specific PIL sentences. In the following analyses was demonstrated the quantitative improvement in readability of sentences that were deemed to need graduate, teacher and researcher levels of skill in reading to be easily understood. It was shown that participants who properly understood the task and were committed could provide valuable revisions even when their reading skill was lower. These can greatly increase the engagement of general audiences and help identify issues that are commonly omitted in the current PPI groups as it was shown the previous chapter that participants with lower level of education are significantly more susceptible to identify issues as severe where more educated participants would classify the same issues as minor even when those issues could make the document incomprehensible to general audiences.

Sample

117 public participants

Sample Size

The target for this study was to recruit 120 participants to revise PIL sentences at graduate, teacher and researcher difficulties. Guidance for sample size in feasibility trials Julious (2005); Sim and Lewis (2012) varies with recommended numbers between 12 to 30+ per arm.

Recruitment

The MTurk crowdsourcing platform was used to recruit 117 public participants who resided in the UK by placing a Human Intelligence Task (HIT) to provide revisions to 27 original PIL sentences.

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.

Results

Participant performance

Each participant was given 9 sentences to revise, Table 6.4 shows the required time to revise a sentence and the time percentiles the participants expended revising each sentence. It was found 95% of all participants (6 fully excluded participants) could help revise at least the easier sentences when they were committed to the task. From those, only 40% (46 participants) could help revise middle difficulty sentences and only 20% (23 participants) with an assessed graduate level reading skill could help improve the most difficult sentences while guarantying the participant had at least a basic understanding of the sentence. Participants with a GSCE reading skill level were found to provide revisions with lower readability scores improving the mean readability of the sentences by 3.5 grades in comparison with 2.5 grades improvement from the A-level group and 2.8 grades for the graduate group.

Table 6.4: Time expended on each sentence per participant.

Participant performance	
GSCE skill level	
n	71
Mean time	0:04:32
Mean readability improvement (US Grades)	3.5
A-Level skill	
n	23
Mean time	0:05:03:78
Mean readability improvement (US Grades)	2.5
Graduate level skill	
n	23
Mean time	0:04:02
Mean readability improvement (US Grades)	2.8
Total number of participants accepted	117

Participant selection

The Cloze procedure was employed to select participants, the participants were presented with sample sentences in which some words were omitted and asked to fill-in the blanks. The participants must be able to correctly fill at least correctly 30% of the spaces for their revisions to be taken into account for sentences of that difficulty.

This prove to be empirically a good cut point between the use of synonyms, grammar mistakes in filling the options and participants not having a clear enough understanding or interest in the task to provide good sentence revisions. With this method the feedback of low educated participants who understood the task and proposed revisions based on employing synonyms to complex words were also included for consideration.

Crowdsourcing the revision of PIL sentences

The second main part of this study assessed the use of crowdsourcing to revise PIL sentences that were deemed to be too hard to be understood by public audiences. A subset of 27 PIL sentences were selected based on their difficulty and topic, three levels of difficulty were established (graduate, post-graduate and professor) and the topics were based on the extended EQIP guidelines. 677 sentence revisions were provided by 117 participants who were recruited via the MTurk platform. Employing crowdsourcing to revise PIL sentences was found to improve the readability of the sentence by at least 1 grade (US school grade) in 70% of the cases ($n_{rev} = 677,\ n_{par} = 117,\ avg_{imp} = 3.2\ median =$

2.5, $percentile_{25} = 0.66$, $percentile_{30} = 0.99$, $percentile_{50} = 2.5$, $percentile_{75} = 4.9$), where n_{rev} is the number of proposed revisions collected, n_{par} , is the number of participants in the study and avg_{imp} is the average improvement in readability of the proposed revision with respect to the original sentence. Figures 6.4 and 6.5 shows the time expended revising the sentences based on their difficulty ranking.

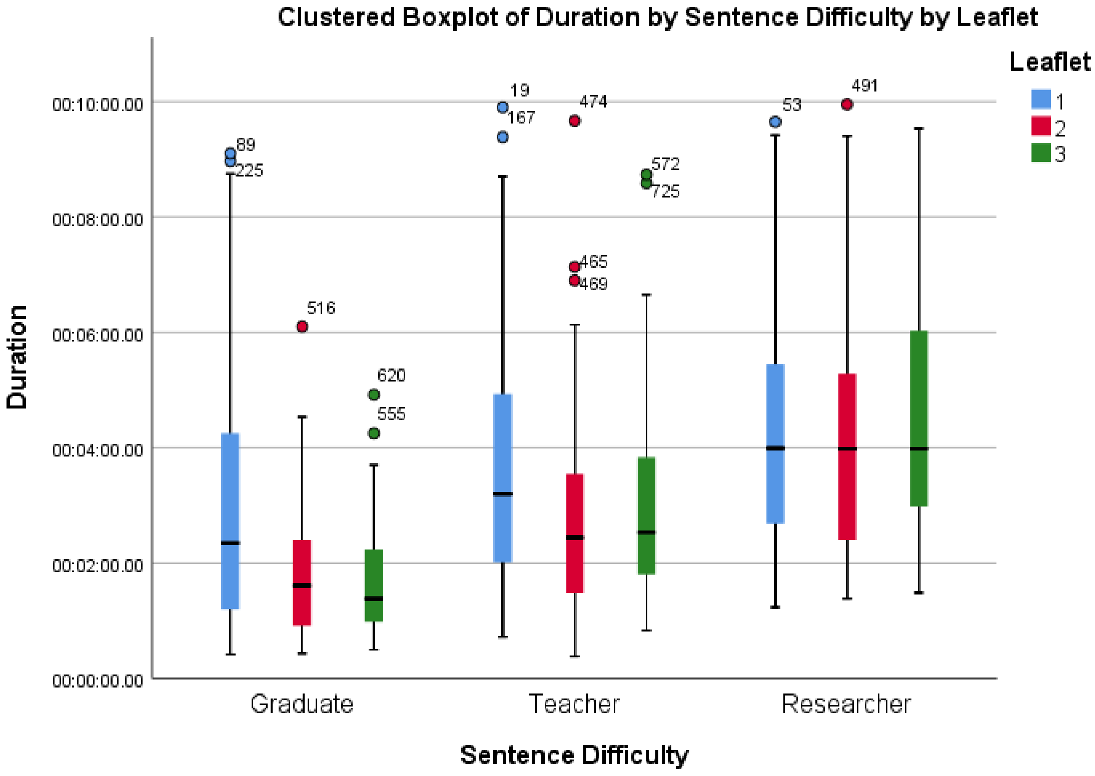


FIGURE 6.4: Time taken to revise a sentence based on the sentence difficulty

Figure 6.4 shows a box plot of the mean times needed by the participants to revise the original sentences. Each selected PIL sentence was catalogued as needing graduate, teacher or researcher reading skill level to be easily understood. Participants were classified in accordance to their reading skill into three groups GCSE (Blue), A-Level (Red), Graduate (Green). It was observed that the time needed to revise a sentence increased as the difficulty of the original PIL sentence raised which can be confirmed in Figure 6.5. In addition, it was observed an increase in the variance of the times needed to revise the sentences linked to participants with lower reading skills and also to the increase in the original PIL sentence difficulty. These results observed the expected behaviour of the system and gave empirical data on how to identify bad revisions based on the expected times of the participants.

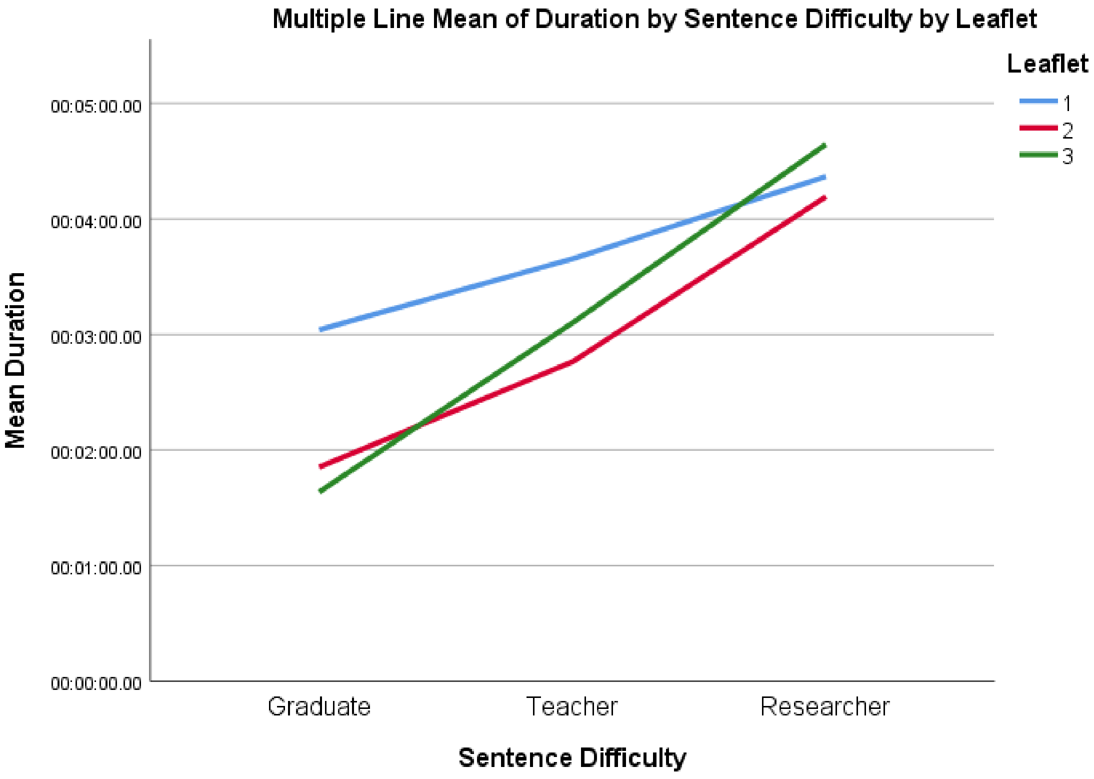


Figure 6.5: Linear association between sentence difficulty and time taken to revise it

A multivariate linear regression model was used to assess the inherent associations of the sentences textual characteristics and with the time needed by the participants to reword them. The independence of the observations was assessed by a Durbin-Watson statistic of 1.816.

A total of 59 extreme outliers were identified and excluded from the analysis, exploration of the data showed this corresponded to cases were the participant interrupted their tasks and concluded them at a later date.

The model showed a statistically significant association between the time needed by the participants to revise a sentence and the sentence difficulty ($p < 0.001$), the number of characters in the sentence ($p < 0.008$) and its readability scores (ARI ($p < 0.001$), Gunning-Fog ($p < 0.016$)), $R^2 = 0.221,\ F(4, 672) = 47.659,\ p < 0.001$. The regression equation found by the model was $Duration = 18.52 + 41.1SentDif + 0.5NumChar + 10.3ARIScore - 5.5GunFogScore$

This model indicates that an association between the readability of the original sentences (given by the readability scores), their complexity (as denoted by the length in characters), the level of improvement in readability (observed as the mean difference between the readability score of the original sentence and its revisions) and the time the participants needed to revise the sentences. These results are consistent with the expected performance of the system.

Readability improvement

The last sections have shown most participants were able to provide revised sentences that were easier to understand than the original sentence. As a final part of this analysis, the factors associated to these readability improvements were searched. It was found the mean improvement on the readability of the proposed revisions was linked to the presence of complex words in the original sentence, the length of the original sentence and the time needed to revise the sentences. This confirms the expectations of easier sentences producing less improvement as the revisions need lower changes.

A multivariate linear regression model was used to assess the inherent associations of the sentences textual characteristics and with the time needed by the participants to reword them. The independence of the observations was assessed by a Durbin-Watson statistic of 1.78.

A total of 59 sentence revisions were excluded from the analysis as exploration of the data showed they corresponded to cases were the participant interrupted their task and concluded them at a later date.

The model showed a statistically significant association between the improvement in the readability of the sentence and the number of complex words and characters in the original sentence, and the time required for the participant to revise the sentence, $R^2 = 0.407,\ F(3, 676) = 47.659,\ p < 0.001$. The regression equation found by the model was $Improvement = -1.89 + .42CompWords - .006Time + .034NumChar$

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

A feasibility study on employing the MTurk crowdsourcing platform to identify the best proposed revisions for PIL sentences that are too hard to be understood by public audiences.

Sample

32 public participants

Sample Size

The target of this study was to obtain 2430 submitted grades for 81 revisions of 9 original sentences. This was considered enough to test the viability of employing an Anova test to identify preferred revisions based on the grades given to each particular revision by the participants. This study was limited in its scope by the restrictions of time and resources available at the time as assessing even a minor sample of all the revisions to a particular sentence would have required hundreds of participants.

Recruitment

Adult participants who resided in the UK were recruited by posting a Human Intelligence Task (HIT) on the UK Amazon Turk platform. A reward of $1 usd was granted for the completion of the task. The participants were selected and grouped based on their education level (GCSE, grad and post-grad).

Procedure

A set of 27 original PIL sentences was selected based on their difficulty and topic. A set of 9 random revision options for each original sentence was presented to each participant. The task consisted of ranking the options in order of preference to replace the original sentence.

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. The participant preference was measured based on the ranking grade given to each option. A one-way Anova model was used to identify significant differences in the preference for a particular revision.

Results

Validation of public revisions' of PIL sentences via crowdsourcing

The final part of this study assessed the use of crowdsourcing to find the best revisions for the original sentences. 2,394 revisions were assessed by 32 participants. Each participant was presented with 9 randomly selected options per sentence for 9 original PIL sentences and asked to rank them. A one-way ANOVA test was used to determine if there were statistically significant differences in the grades given by the participants between a set of 9 proposed revisions for sentences that were deemed too hard to be understood by general audiences. Only in too cases were found statistically significant differences, ($F(26, 505) = 2.766$ *and* $F(26, 505) = 2.637$ *both with* $p < .001$). A Tukey post hoc test revealed that the grades given to these proposed revisions were statistically significantly less (indicating a higher preference for those options) ($.37 \pm 2.753,\ p < .001$) and ($.89 \pm 3.381,\ p < .001$) than other options.

Chapter 7

Discussion and Conclusions

7.1 Summary of findings

This section highlights the main findings of the research project, it is divided in three main areas: the analysis of PIL textual characteristics, the analysis of public feedback to PILs and the analysis of the effect on sentence readability of employing crowdsourcing, and content analysis techniques to identify, revise and validate readability issues through a Webtool.

7.1.1 PIL characteristics (Chapter 4)

The first objective of this thesis was to understand the association of various textual characteristics of the PILs and the performance of clinical trials. This correlational study analysed 58 NIHR funded trials with publicly available PILs. The performance of each PIL was measured based on the percentage of the planned sample actually recruited by the trial. The trial information was provided by the NETSCC and included the type of trial, the clinical area researched, the sample size, project dates, trial duration, setting, number of recruitment centres, and the trial ICD_{10} and HRCS codes.

Three statistically significant models were found in this study, showing strong associations between the Coleman-Liau ($R^2 = 0.86,\ p = 0.026$) SMOG ($R^2 = .876,\ p < 0.001$) and Flesch-Kincaid ($R^2 = 82.6,\ p < 0.001$) readability index scores with the percentage of the planned sample recruited by the trials. The first model was a univariate linear regression between the percentage recruited and the Coleman-Liau score. The second model was a multivariate linear regression model for the SMOG index, which included associations with there being less than 10 recruitment centres, being a treatment evaluation trial and having a diagnostic intervention. The final model describes a multivariate linear regression model for the recruited percentage of the planned sample that included associations with the Flesch-Kincaid score, the number of words associated with Joy present in the document, the sample size being increased during the study, being a drug intervention and the number of words associated with Negative sentiments. This is the first time that such strong associations have been demonstrated between PIL readability scores and trial recruitment rates, suggesting that PIL content does influence study recruitment rates. An alternative hypothesis being that carefully designed trials would focus more attention to the quality of all trial aspects, which would be reflected in higher overall performance.

7.1.2 Public feedback on PILs (Chapter 5)

The main objective of the second study was to assess the characteristics of public feedback on PILs. This descriptive study confirmed significant differences between both understanding of key PIL content (such as the risks that trial participants face) and overall quality perception of the PIL text based on the participants' education level (GCSEs, Undergrad and Graduate). The results confirmed that understanding of the key information in PILs increased with the level of education ($R^2 = 0.75,\ p < 0.01$). The percentage of correct answers to questions about key trial details given by the GCSE educated group did not vary significantly from their expected accuracy if they choose their answers randomly. Also, no significant association was found between the participants perception of PIL quality and their understanding of the key trial information, for all of the groups.

Thematic analysis of comments focused on specific parts of PILs and those referring to the overall document showed different roles, themes, and amounts of emotion. General comments were found to be more emotive and express commendations and endorsements or directly confront the author. On the other hand, specific comments contained more requests for change or further explanation or pointed out mistakes in the information in a neutral tone.

The thematic analysis showed that 86% (128 of 149 comments) of the topics approached by the participants in their specific comments corresponded with the topics listed in the Expanded EQIP scale, which is a validated instrument to assess the quality of information intended for patients. Other topics mentioned by the participants that were not directly present in the scale were: the consent and randomization processes, the presence of redundant information, and bad writing, including grammar, spelling, and punctuation mistakes or the use of inappropriate language. Looking at correlations between participant comments and understanding (measured as the percentage of correct answers they gave), the only correlation that was statistically significant was the number of specific comments given by a participant ($R^2 = .196,\ p < 0.01$).

7.1.3 Employing content analysis, readability indexes and crowdsourcing to improve the readability of PILs (Chapter 6)

PIL author feedback on the proposed Webtool

The feedback from a focus group of 5 PIL authors made it clear that a Webtool to help them revise PILs would be of value to them. This Webtool had the capability to collect public feedback via a website, permit the assessment of PIL quality and understanding of key trial details through questionnaires based on the extended EQIP guidelines, calculate the document readability metrics, highlight sentences with readability issues, and present visual reports of this data via a website for PIL authors. The PIL authors ranked the visualization of comments related to each section of the PIL as the most useful feature of the platform. They expressed their belief that some of the EQIP questions would not be appropriate for all PILs, and that it was necessary for the Webtool to give advice on how to solve the readability issues after identifying them.

Identifying PIL sentences that are too hard to be readily understood

Five different readability indexes were used to assess the readability of the PILs in two studies. The first study, assessed the overall readability of 58 documents and found no significant differences were found in the readability scores given by these indexes to the full text of the PILs. In our third study we employed the readability scores to identify particular sentences that were too hard to be widely understood by public audiences for 3 PILs in current use by considering the degree of agreement between the readability scores given to each sentence. All PILs were found to have overall readability scores above 10th grade in all the indexes and an average of 34.2% of the sentences being deemed to be too hard and a standard deviation of 2.5. Sentences with more than 15 words and for which at least three of the indexes provided scores above 9th grade, which is the maximum reading difficulty score suggested by the literature for information that needs near universal understandability.

Crowdsourcing the revision of PIL sentences

The third study recruited 117 public participants via MTurk to obtain 677 revisions for 27 original PIL sentences that were deemed to be too hard to be easily understood. Significant associations were found between the number of complex words and characters in the original sentence and the readability improvement of the proposed revisions. The readability improvement was also found to be negatively associated with the time the participants expended revising the sentence, ($R^2 = .407,\ p < 0.001$).

The average time expended by the participants reviewing the sentences was 4:33 minutes per sentence with interquartile range of $[Q_1 = 2 : 39,\ Q_3 = 5 : 50]$ minutes. The time taken was found to be associated with the sentence difficulty, the number of characters in the original sentence, and two readability scores (ARI, Gunning-Fog).

Validation of the proposed revisions of PIL sentences

A random sample of 9 proposed revisions for each of 9 PIL sentences was selected to be validated via crowdsourcing. 32 people viewed a set of 266 proposed revisions and used a unipolar scale to assess their readability. A one-way ANOVA found statistically significant preference for only two of the proposed revisions based on the grades given by the participants. This shows the participants did not show a clear preference between most of the proposed revisions for the original sentences. A hypothesis can be made that most of the proposed revisions are too similar with only one revision cycle as most participants may have focused on addressing only one issue per sentence.

7.2 Research limitations

Specific limitations of each study have been described in the corresponding chapter, this sections focuses on highlighting those limitations that are particularly important for future research in the area.

First, the analysis of public feedback on PILs was focused on a limited set of 3 PILs that presented the most readability issues from the previous study. This was done to ensure the largest coverage of PIL readability issues when assessing the participants' perception of PIL quality and studying the association between the essential topics on trials proposed by the literature with the topics approached by the participants' feedback. Further studies of the participants' response to more diverse PILs will be required to comprehensively understand the associations between PIL readability, comprehension and perception of quality.

Second, the results obtained from the study on employing content analysis and crowdsourcing via a Webtool to identify, revise and validate readability issues employed several sub-studies. When public participants were required for a sub-study they were recruited via the Amazon Mechanical Turk (MTurk) platform. Further analysis of the data shows that the time and day the tasks are published on MTurk may affect how fast participants are recruited and should be explored in future research. In addition, the increase in skill as MTurk participants engage in revision tasks and their effects on their performance, perception of quality and comprehension level must be added to their current statistical models.

7.3 Research implications

This section focuses on the implications of the results of the empirical studies present in this research. The descriptive study on the textual characteristics of PILs corroborated the initial hypothesis that readability of the PILs was correlated with the recruitment rates to RCTs by finding significant associations between readability scores of the SMOG, Coleman-Liau and Flesch-Kincaid indexes given to each PIL and the percentage of the sample that was recruited to the corresponding RCTs. This suggest that identifying and correcting readability issues in PIL documents could increase the overall recruitment to trials, but a rigorous randomised study within a trial (a SWAT) would be needed to confirm this.

The thematic and content analysis of the public feedback to PILs with serious readability issues confirm the literature findings that the education level of the participants has a significant impact on their understanding and perception of the PIL quality. In addition, it showed there is no significant correlation between the participants' subjective perception of quality and their understanding of the information. Quality perception diverges between participants of different education levels across all PILs with only 14% of the GCSE participants believing the PILs presented minor or no issues in contrast to 68% for undergrad participants and 71% for graduates.

Public comments on PILs were found to be associated with good understanding of PIL content only when they applied to specific sections. This implies the need to incorporate alternative systems or procedures to assess the readability of the PILs instead of the public feedback obtained from Patient and Public Involvement (PPI) groups, as the current definition of public involvement does not permit the direct assessment of how much information was understood by the participants.

The PIL author feedback unanimously agreed that there are readability issues in the current PILs that affect the participants' decision to join RCTs. They also believed that not only is it necessary to identify these issues, but also to propose solutions to the authors to increase the quality of the PIL documents. Several descriptive studies demonstrated that employing content analysis metrics, readability indexes and crowdsourcing can identify, revise and validate sentences that are too difficult to be understood by public audiences. Employing crowdsourcing to revise PIL sentences was found to improve the readability of the sentence by an average of 3.2 grades (US school grade) with a median of 2.5 grades and an interquartile range of [0.67,4.9]. This implies employing crowdsourcing to revise PIL sentences can bring appreciable benefits in improving the readability of the sentences in 75% of the cases and it should be considered by PIL authors in addition to PPI groups when seeking to improve PIL readability. There was significant variation in the readability improvement of the proposed revisions. 41% of this variance was explained by the number of complex words present in the sentence, the number of characters and the amount of time the participant expended in the revision ($R^2 = 0.41,\ p < 0.001$).

The following sections give consideration to the specific impact this research project could have on the core stakeholders.

Trialists

A large impact of this project would be represented by the future application of the proposed methodology to design and ensure PILs to low risk trials are easily understood by their audiences. One of the effects of the current proportionate approach to seeking consent is that documents for trials with low risk to participants face less oversight, which may explain findings of Chapter 4 were it was found associations between the readability of the documents and the number of recruiting centres, the type of trial and the presence of emotive words.

Policy makers

Many independent works have reported that the documents used to inform participants contain severe readability issues that may have made them unfit for purpose as discussed in Section 2.1. Creating a system that helps quantify the impact of the revision process on the readability of the documents would be by itself a valuable tool for policy makers. In addition, this tool could help identify associations between document readability and trial recruitment and retention rates.

Practitioners

One of the key issues of medical practitioners when inviting participants to clinical trials is to maintain equipoise Elliott et al. (2018), where practitioners find difficult to approach all potential participants and held preconceptions on the effect of the treatments. This biases is commonly transmitted through wording, poorly balanced information or direct recommendation to the potential participants Elliott et al. (2018).

Creating documents that are easily understood by all participants could help alleviate these issues by presenting a more measured balance of the essential topics.

Further research

Implementing this research as a national linked database could facilitate the identification of solutions to readability issues by cross-referencing the proposed revisions to similar sentences and terminology with the change made by the researcher and the impact on the documents readability and the trials recruitment and retention rates.

In addition, a similar methodology to receive public feedback on the practitioners presentation of the topics can be explored in future research, by presenting a video of a practitioner to MTurk participants and obtaining their feedback, assessing their understanding of the information and analysing their preferences.

7.4 Future work

Based on the previous considerations, future work to expand this research project should include a comparative analysis of the use of crowdsourcing to revise PIL sentences from a diverse group of PILs, taking as factors the day and time the task is submitted in the MTurk platform, the RCT type and research area, the overall readability of the PIL document and the literacy skill of the participants. In addition, further analysis on the effect of emotive text (or its lack thereof) in the reader capacity to understand and remember PIL information when in an emotive state is needed. An study where the PIL authors assess the best revisions identified by the participants based on their content, appropriateness and quality should be added. Finally, a study to compare the original documents to revised PILs based on the proposed revisions of PIL sentences would permit quantify the effect on the readability of the complete document.

7.5 Discussion

This research project has shown the use of crowdsourcing can significantly improve the readability of leaflets intended to inform potential participants in randomized controlled trials. One of the key issues identified in the current proportionate approach to seek informed consent based on the risk to the patient is that documents intended for trials with lower risks commonly pass oversight without needing to ensure they are understandable to the intended audience. Making an objective measurement of the understandability of a document is also an issue with the current division of Public Involvement vs research, where the PIL authors can not assess the level of understanding of the reader when working with a Patient and Public Involvement group or directly use participant data as part of the research as this would make them fall outside the scope of PPI.

Crowdsourcing is thus a valuable tool to provide the essential information the PIL authors need not only to obtain revisions of specific sentences that could also be obtained with a PPI group but to quantify the effect the revision process had in the readability of the document final version. This is not to say this proposal sought to fully replace the human element in the process of deciding how to correct the PIL issues. The proposed framework just present an alternative way to collect feedback from participants while using a Webtool in a manner that helps quantify the readability of the original version, keep track of readability issues and the provenance of potential solutions. The authors will receive the participants' feedback, the visualization of the readability issues identified by the system with readability indexes and the preferred proposed revisions made and selected by public participants. Then, they will make the corrections base on the professional knowledge and facilitated information. The system can then also quantify the readability of the final version to help understand the impact this process had on the PIL readability. Based on the empirical data of this project, this process could take as few as 10 days excluding the time needed for the researcher to revise the document.

7.6 Conclusions

During the last two decades, several studies have consistently found readability issues in PILs intended to inform participants who are being recruited to RCTs. Solving these issues has become a priority area of research for UK organizations linked to health research and have brought forward the current model of proportionate approach to seeking consent. The Table 7.1 presents a comparison between the current model and the proposed framework to design PIL with better readability.

The main objective of the first study was to assess the readability and characterize PILs supported by the HRA with publicly available documents. This research project focused on identifying associations between trial recruitment rates and PIL document characteristics. A linear analysis showed significant correlation between the readability score (Smog index) and the recruited proportion of the planned sample, having 10 or less centres to recruit participants, and being a trial for evaluating a treatment. It was also found significant associations between the recruited proportion of the planned sample and the Flesch-Kincaid reading score, the amount of words related to joy and negative emotions, having the sample size increased and having a drug intervention. When grouped by type or research area the PILs demonstrated large variances in their number of words, characters, syllable, sentences and complex words, but all the readability indexes gave consistent measurements on the education level needed to understand the PILs in all cases requiring more than 10th grade education and thus being above the literature recommended maximum limit to be easily readable by general audiences.

A comparison of the emotion present in the PILs with news articles from the BBC, Daily Mail and Hello Magazine to anchor the scale with a gradient of texts intended to inform different segments of public audiences. Here the results shown that the PILs contained significantly more words related to positive emotions and significantly less related to negative emotions than any of the other groups. On the other hand, the PILs included significantly less words related to other emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) assessed. These results shown that the readability of PILs and their proportion of emotive content can be associated with the recruitment rates on their trials. The heavy skew of positive to negative words and lack of words related to other emotions could be a result of the current guidance on maintaining a respectful tone of voice while engaging the patient, but literature on emotion and behaviour show that readers in emotive states may find it difficult to pay attention, understand and remember information which lacks emotive stimulus or that greatly diverges from their emotive state. This implies a need to further research the effect of emotive text (or lack thereof) on information intended for patients who maybe in an emotional state.

The main focus of the second study was to explore the associations between participant understanding and their perception of information quality in public audiences. The literature in the area has consistently found severe issues with the readability of PILs for recruiting participants into RCTs, and that most RCTs struggle to successfully recruit at least 80% of their planned samples. Even though this has been a priority area of research for UK health research organizations in the last two decades, the study confirmed the issues expressed in the literature on current PIL leaflets and explored a novel approach to understanding these issues.

Three groups of public participants were studied based on their education level (GCSEs, Undergrad, Grad) who assessed 4 PILs with significant readability issues based on four readability indexes (ARI, Gunning-Fog, Smog, and Flesch-Kincaid). No significant association was found between their qualitative perception of the quality of the PILs information and their understanding of the information for any of the groups. The groups presented distinct impressions of the PILs quality in all cases, but the Undergrad and Graduate groups had significantly more participants who only recognized minor o no quality issues with the leaflets (15% against 68% and 71% respectively). This means a simple assessment of the PIL where a group of recruited participants use a scale to determine if it can be used is an unreliable method to measure the quality of RCT PILs.

The thematic analysis of the participants' feedback showed only comments associated to specific sections of the PILs were correlated to the participants' understanding of the information. In addition, the intentions, topics and emotions of general and specific comments were found to be significantly different and the comments given by the participants were found to closely follow the identified topics by EQIP guidelines to assess information quality. This thematic analysis also showed that employing focus groups to review the PILs intended to recruit participants into RCTs can help identify issues in the documents, but the process must be guided to focus the participants on giving specific feedback instead of emotive context. Furthermore, the analysis shown some areas that are not currently not directly assessed by the EQIP scale like the clarity of the randomization process, the inclusion of repeated or redundant information, and the presence of bad writing (grammar, spelling, and punctuation mistakes or the use of inappropriate terms).

These results have generated insights on some aspects of the current HRA model of “proportionate approach to seeking consent”, which seeks to separate clinical research and public involvement, where public involvement is defined as “research being carried with or by members of the public rather than to, about or for them”. Given that the current HRA guideline to address PIL information quality issues is to engage a Patient and Public Involvement (PPI) group, this definition discourages the researcher from assessing the participants’ understanding of the PIL information, and instead asks them to rely directly on the participants’ perception of quality. As the study shows, these assessments do not reliably reflect the PIL’s capacity to inform the participants about key issues they need to consider when deciding whether to join the trial.

The descriptive studies on employing readability indexes and content analysis to identify and visualize issues in the PILs were found to be of great value by the PIL authors, in addition their feedback indicate that it is necessary to provide proposals on how to resolve these issues. Several descriptive studies analysed the use of crowdsourcing to obtain and validate proposed revisions. It was found that in at least 70% of the cases the readability of the sentence improved by 1 grade or more; sentences that were harder to read showed greater improvements. The validation of the proposed revisions via crowdsourcing only found statistically significant differences in the grades given to revisions in two cases, which shows the participants did not have a clear preference for any particular revision in most cases. This implies the need to employ other validation methods or receive expert feedback on the revisions.

In conclusion, this research project has found current PIL documents still contain readability issues consistent with findings reported in the literature review; that there is a correlation between readability scores and the recruitment rates to RCTs; that readability indexes can consistently identify readability issues in PILs with large divergence in their lexical characteristics; that the public feedback on PILs follows the expected topics proposed by the literature to assess information quality; that exist significant differences in the participants’ perception of the issues in the documents depending on their education level; and that employing crowdsourcing to revise PIL sentences can improve their readability by at least a grade in 70% of the occasions.

Table 7.1: Comparison between the current PIL development model and the proposed framework for improving PIL readability by engaging members of the public through a crowd sourcing platform

Model	Advantages	Disadvantages

Continued on next page

Table 7.1 – continued from previous page

Model	Advantages	Disadvantages
Current model of public involvement: 1. Draft a PIL based on the NHS template 2. Engage patients or members of the public to review the PIL following the INVOLVE guidelines (optional for most trials) 3. Revise the PIL based on the public feedback obtained 4. Present the PIL for consideration as part of the ethical review process needed to approve a trial	1. Helps identify research priorities from a patient perspective 2. Highlights diverse patients needs for specific groups 3. Offers flexibility of document design based on the risk to the patients	1. No standardized process: activities can range from full collaboration to brief consultation or even token participation of public collaborators 2. Presents difficulty in defining the impact of public involvement 3. Flexible regulations commonly lead to inadequate reporting 4. Lacks reliable and valid tools to assess the impact of public involvement on the PIL 5. The high-cost of public engagement defined in current guidelines (£25 per hour per person) make it infeasible for low risk, unfunded or pragmatic trials

Continued on next page

Table 7.1 – continued from previous page

Model	Advantages	Disadvantages
Proposed model: 1. Develop a PIL based on the NHS template 2. Assess the readability of the PIL using quantitative metrics 3. Engage members of the public through crowdsourcing 4. Revise the PIL based on the gathered feedback 5. Reassess new PIL version with public reviewers or a sample of the target population 6. Present the PIL for consideration as part of the ethical review process needed to approve a trial	1. Permits to quantitatively assess the effect of engaging members of the public based on readability scores and the percentage of correct answers given by the participants 2. Lowers the cost of engaging members of the public enough to make it feasible for all studies 3. Permits us to identify associations between comment topics, PIL readability, reading skill, and participant understanding	1. Under the current definition used by INVOLVE, it may not be considered ”public involvement” 2. The new process requires the development of a Webtool to facilitate the gathering and processing of the data by the trialist

Continued on next page

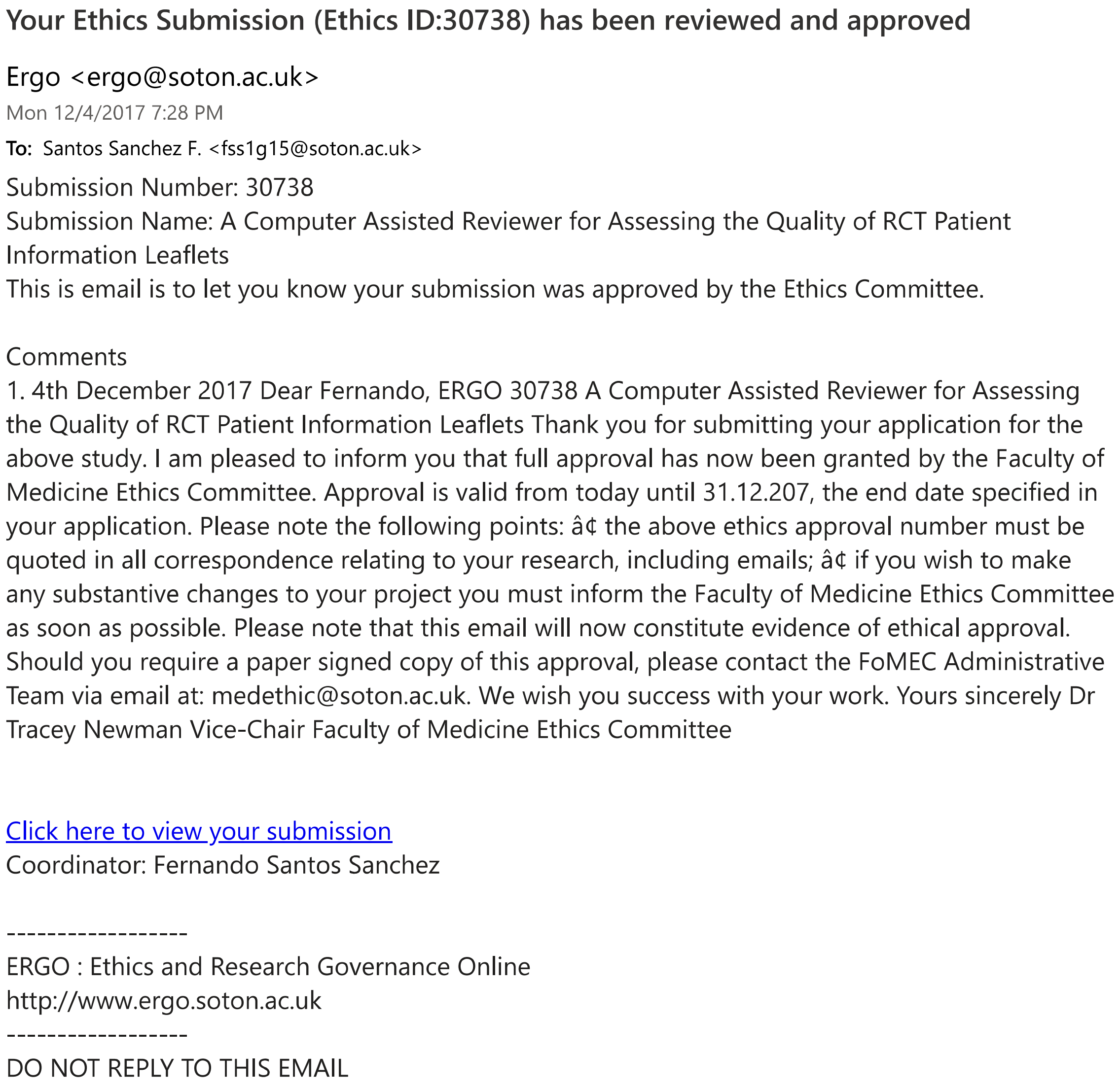
Table 7.1 – continued from previous page

Model	**Advantages**	**Disadvantages**
Proposed model: 1. Develop a PIL based on the NHS template 2. Assess the readability of the PIL using quantitative metrics 3. Engage members of the public through crowdsourcing to: a) Propose revisions for sentences that are too hard to be widely understood b) Read and comment on the PIL information c) Assess the reader skill through quantitative methods like Cloze procedure d) Assess reader understanding of the key trial points through questionnaires based on EQIP guidelines e) Use crowd members to help identify the best proposed revisions for each sentence 4. Revise PIL based in the feedback 5. Reassess the new PIL version 6. Present PIL for consideration	1. Permits to quantitatively assess the effect of engaging members of the public based on readability scores and the percentage of correct answers given by the participants 2. Lowers the cost of engaging members of the public enough to make it feasible for all studies 3. Permits us to identify associations between comment topics, PIL readability, reading skill, and participant understanding	1. Under the current definition used by INVOLVE, it may not be considered "public involvement" 2. The new process requires the development of a Webtool to facilitate the gathering and processing of the data by the trialist

Appendix A

Ethical approval emails

A.1 Study I



Your Ethics Submission (Ethics ID:30738) has been reviewed and approved

Ergo <ergo@soton.ac.uk>
Mon 12/4/2017 7:28 PM
To: Santos Sanchez F. <fss1g15@soton.ac.uk>

Submission Number: 30738
Submission Name: A Computer Assisted Reviewer for Assessing the Quality of RCT Patient Information Leaflets
This is email is to let you know your submission was approved by the Ethics Committee.

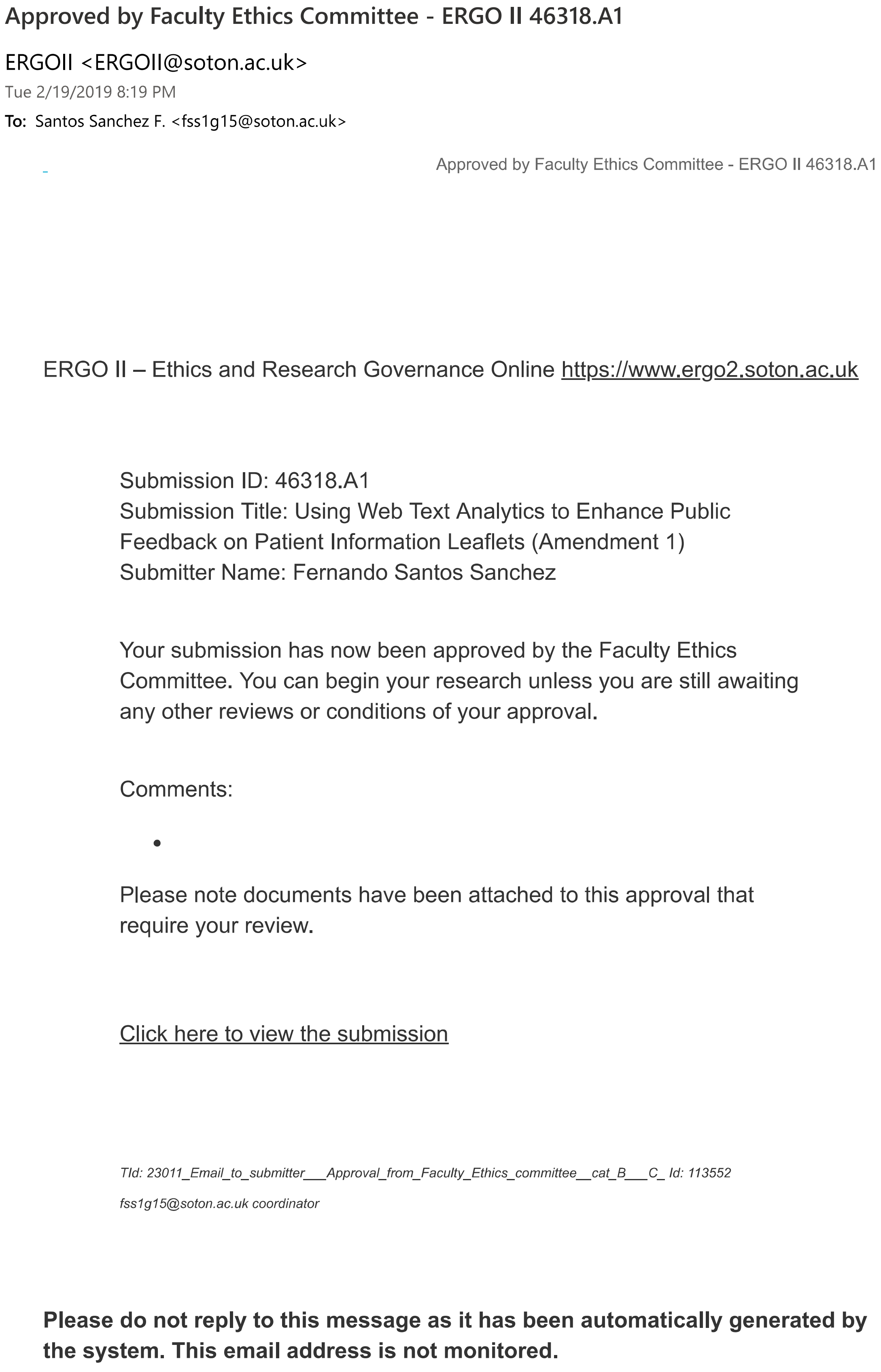
Comments
1. 4th December 2017 Dear Fernando, ERGO 30738 A Computer Assisted Reviewer for Assessing the Quality of RCT Patient Information Leaflets Thank you for submitting your application for the above study. I am pleased to inform you that full approval has now been granted by the Faculty of Medicine Ethics Committee. Approval is valid from today until 31.12.207, the end date specified in your application. Please note the following points: â¢ the above ethics approval number must be quoted in all correspondence relating to your research, including emails; â¢ if you wish to make any substantive changes to your project you must inform the Faculty of Medicine Ethics Committee as soon as possible. Please note that this email will now constitute evidence of ethical approval. Should you require a paper signed copy of this approval, please contact the FoMEC Administrative Team via email at: medethic@soton.ac.uk. We wish you success with your work. Yours sincerely Dr Tracey Newman Vice-Chair Faculty of Medicine Ethics Committee

Click here to view your submission
Coordinator: Fernando Santos Sanchez

ERGO : Ethics and Research Governance Online
http://www.ergo.soton.ac.uk

DO NOT REPLY TO THIS EMAIL

A.2 Study II



Approved by Faculty Ethics Committee - ERGO II 46318.A1

ERGOII <ERGOII@soton.ac.uk>
Tue 2/19/2019 8:19 PM
To: Santos Sanchez F. <fss1g15@soton.ac.uk>

Approved by Faculty Ethics Committee - ERGO II 46318.A1

ERGO II – Ethics and Research Governance Online https://www.ergo2.soton.ac.uk

Submission ID: 46318.A1
Submission Title: Using Web Text Analytics to Enhance Public Feedback on Patient Information Leaflets (Amendment 1)
Submitter Name: Fernando Santos Sanchez

Your submission has now been approved by the Faculty Ethics Committee. You can begin your research unless you are still awaiting any other reviews or conditions of your approval.

Comments:

-

Please note documents have been attached to this approval that require your review.

Click here to view the submission

TId: 23011_Email_to_submitter___Approval_from_Faculty_Ethics_committee__cat_B___C_ Id: 113552

fss1g15@soton.ac.uk coordinator

Please do not reply to this message as it has been automatically generated by the system. This email address is not monitored.

Appendix B

Recruitment materials

B.1 Study I

Public Involvement

Are you interested in making a difference?
Do you believe the information given to patient could be better?

Join Us!

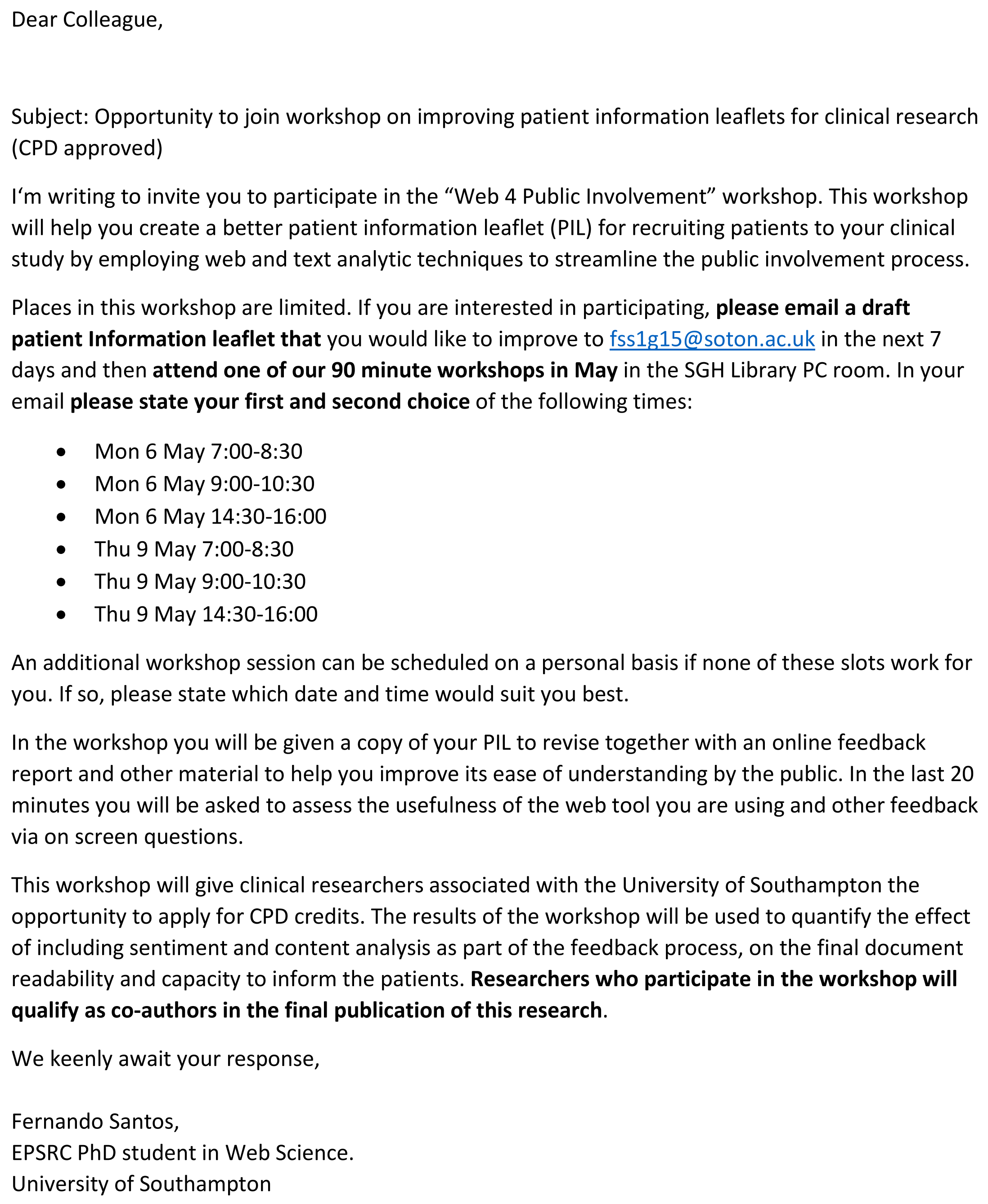
Learn how to give high-quality feedback and comments on the information given to patients of previous low-risk trials in the UK and assess how well the information was presented be part of making better information resources.

Win up to £20 for your participation!
Register Online at:
https://fercom2000.wixsite.com/carpi

CARPI
Public Involvement Group

UNIVERSITY OF
Southampton

B.2 Study II



Dear Colleague,

Subject: Opportunity to join workshop on improving patient information leaflets for clinical research (CPD approved)

I‘m writing to invite you to participate in the “Web 4 Public Involvement” workshop. This workshop will help you create a better patient information leaflet (PIL) for recruiting patients to your clinical study by employing web and text analytic techniques to streamline the public involvement process.

Places in this workshop are limited. If you are interested in participating, **please email a draft patient Information leaflet that** you would like to improve to fss1g15@soton.ac.uk in the next 7 days and then **attend one of our 90 minute workshops in May** in the SGH Library PC room. In your email **please state your first and second choice** of the following times:

- Mon 6 May 7:00-8:30
- Mon 6 May 9:00-10:30
- Mon 6 May 14:30-16:00
- Thu 9 May 7:00-8:30
- Thu 9 May 9:00-10:30
- Thu 9 May 14:30-16:00

An additional workshop session can be scheduled on a personal basis if none of these slots work for you. If so, please state which date and time would suit you best.

In the workshop you will be given a copy of your PIL to revise together with an online feedback report and other material to help you improve its ease of understanding by the public. In the last 20 minutes you will be asked to assess the usefulness of the web tool you are using and other feedback via on screen questions.

This workshop will give clinical researchers associated with the University of Southampton the opportunity to apply for CPD credits. The results of the workshop will be used to quantify the effect of including sentiment and content analysis as part of the feedback process, on the final document readability and capacity to inform the patients. **Researchers who participate in the workshop will qualify as co-authors in the final publication of this research**.

We keenly await your response,

Fernando Santos,
EPSRC PhD student in Web Science.
University of Southampton

Appendix C

Research Data

C.1 Literature Review

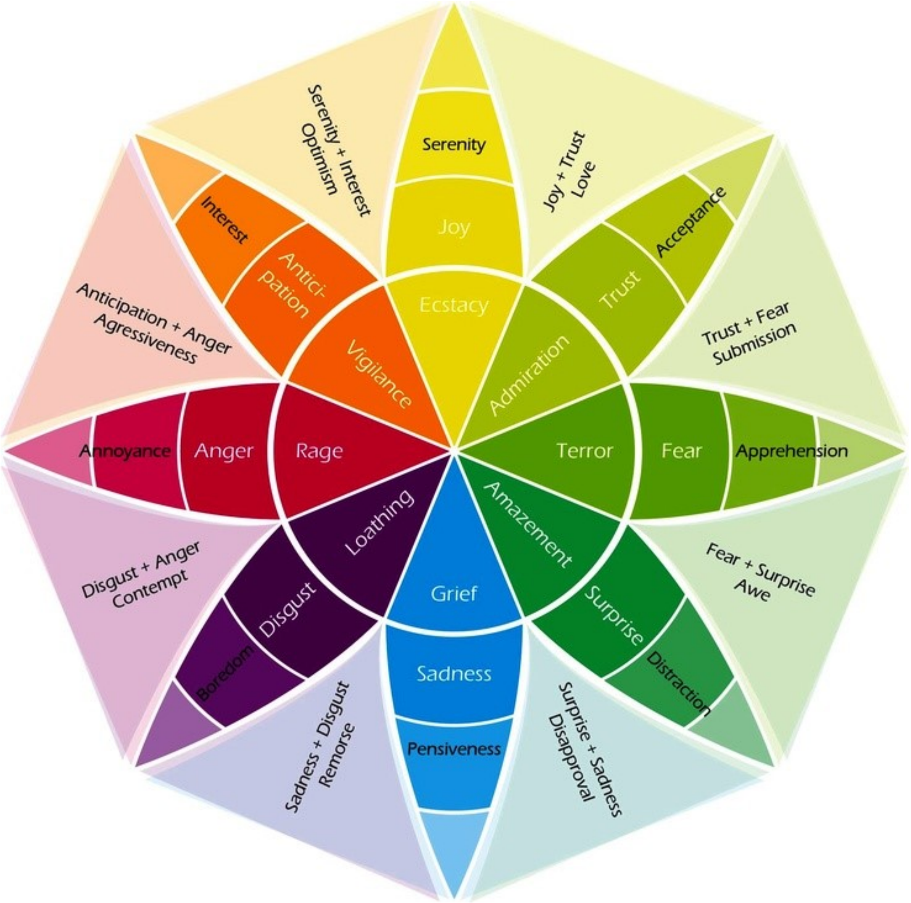


FIGURE C.1: Plutchik's Wheel of Emotion Plutchik (1984)

C.2 Study I

Table C.1: PIL characteristics.

PIL id	Words	Characters	Syllables	Sentences	Complex words
p6	1348	6239	2068	76	175
p7	1437	6482	2104	89	141
p11	1867	8522	2780	88	202
p12	413	1967	643	18	59
p13	1704	8022	2599	85	236
p14	1572	7041	2292	70	177
p18	522	2409	785	19	66
p19	1273	5819	1966	61	176
p23	817	3821	1242	45	110
p25	1861	8169	2709	93	183
p27	1791	8449	2742	99	251
p28	2030	9369	3067	130	240
p29	1784	8487	2820	91	277
p31	1271	6168	2086	75	228
p33	2184	9992	3298	107	261
p35	4915	22608	7437	212	604
p37	1759	8937	3028	88	351
p38	5849	25991	8526	235	608
p39	1419	6842	2249	69	191
p40	2706	12207	3996	144	280
p41	1900	8831	2884	97	230
p47	871	3990	1356	46	110
p49	947	4481	1426	50	118
p50	2266	11143	3787	153	382
p52	1762	7988	2669	94	214
p56	1858	9076	2936	71	263
p64	3066	14358	4752	127	407
p68	522	2651	888	19	101
p71	463	2164	710	22	63
p75	701	3366	1128	35	117
p79	624	3025	1052	22	111
p81	586	2491	808	22	41
p89	1755	7960	2632	112	205
p90	5552	25884	8436	291	765
p93	1358	6227	2064	57	163
p96	775	3644	1204	32	107
p100	1049	5127	1698	51	156

Continued on next page

Table C.1 – continued from previous page

PIL id	Words	Characters	Syllables	Sentences	Complex words
p101	2087	9463	3152	102	252
p102	1030	4603	1538	62	135
p104	1220	5630	1871	64	144
p108	517	2345	759	16	57
p112	1089	4898	1571	78	99
p114	664	2948	958	24	65
p117	759	3488	1142	41	92
p118	1520	7288	2433	74	237
p119	966	4464	1473	40	128
p120	1262	5994	1911	58	151
p121	874	4030	1328	48	118
p122	1693	8030	2685	78	234
p125	989	4515	1488	57	112
p129	846	3962	1304	44	120
p131	565	2559	830	22	57
p132	878	3861	1297	41	97
p134	2848	12431	4109	146	311
p138	2119	9546	3068	109	222
p140	1551	7123	2377	85	203
p142	2364	10673	3450	115	260
p144	1008	4738	1582	43	139

Bibliography

Representativeness and diversity of people who get involved. 2020.

Mery Natali Silva Abreu, Arminda Lucia Siqueira, Clareci Silva Cardoso, and Waleska Teixeira Caiaffa. Ordinal logistic regression models: application in quality of life studies. *Cadernos de Saúde Pública*, 24:s581–s591, 2008.

Muhammed O Afolabi, Joseph U Okebe, Nuala Mcgrath, Heidi J Larson, Kalifa Bojang, and Daniel Chandramohan. Informed consent comprehension in a frican research settings. *Tropical medicine & international health*, 19(6):625–642, 2014.

Chanel Agness, Erica Murrell, Nancy Nkansah, and Caren McHenry. Poor health literacy as a barrier to patient care. *The Consultant Pharmacist®*, 23(5):378–386, 2008.

Tanja Aitamurto, Aija Leiponen, and Richard Tee. The promise of idea crowdsourcing–benefits, contexts, limitations. *Nokia Ideasproject White Paper*, 1:1–30, 2011.

Amazon. Amazon mechanical turk, faq page, 2017.

Cande V Ananth and David G Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6): 1323–1333, 1997.

Dimitra Anastasiou and Rajat Gupta. Comparison of crowdsourcing translation with machine translation. *Journal of Information Science*, 37(6):637–659, 2011.

W.A. Association. Web analytics definitions, 2007.

BBC-Bitesize. Target audience, 2018.

Emma Bell, Alan Bryman, and Bill Harley. *Business research methods*. Oxford university press, 2018.

Bernard Berelson. Content analysis in communication research. 1952.

Irma Borst, Christine Moser, and Julie Ferguson. From friendfunding to crowdfunding: Relevance of relationships, social media, and platform activities to crowdfunding performance. *new media & society*, 20(4):1396–1414, 2018.

Daren C Brabham. *Crowdsourcing*. Mit Press, 2013a.

Daren C Brabham. *Using crowdsourcing in government.* IBM Center for the Business of Government, 2013b.

Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.

Brittany Canfield. Health literacy universal precautions: A quality improvement project to promote effective use of clear, plain language communication within primary care. 2020.

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

F.E. Clements. Use of cluster analysis with anthropological data. *American Anthropologist*, 56:180–199, 1954.

Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

Mike Conway. The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, 83 (1):186–200, 2006.

Julie M Rudgers Croft. *The Readability of Pediatric Dentistry Patient Education Materials.* PhD thesis, University of Illinois at Chicago, 2012.

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

C. Darwin. The expression of the emotions in man and animals, 1998.

C. Darwin, M.J. Adler, and R.M. Hutchins. The origin of species by means of natural selection, 1872.

Thomas Dobbs, Giles Neal, Hayley A Hutchings, Iain S Whitaker, and James Milton. The readability of online patient resources for skin cancer treatment. *Oncology and Therapy*, 5(2):149–160, 2017.

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

Daisy Elliott, Freddie C Hamdy, Tom A Leslie, Derek Rosario, Tim Dudderidge, Richard Hindley, Mark Emberton, Simon Brewster, Prasanna Sooriakumaran, James WF Catto, et al. Overcoming difficulties with equipoise to enable recruitment to a randomised controlled trial of partial ablation vs radical prostatectomy for unilateral localised prostate cancer. *BJU international*, 122(6):970, 2018.

Lubna Elmadani. Readability and suitability of online noise-induced hearing loss information in english. 2019.

M.J. Escudero-Carretero, S. Sánchez-Gómez, R. González-Pérez, R. Sanz-Amores, M.A. Prieto-Rodríguez, and E. Fernández de la Mota. Elaboración y validación de un documento informativo sobre adeno-amigdalectomıa para pacientes. *Anales del sistema sanitario de Navarra*, 36:21–33, 2013.

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

National Institute for Health Research NHS. What is public involvement in research?, 2018.

J Francisca Caron-Flinterman, Jacqueline EW Broerse, Julia Teerling, and Joske FG Bunders. Patients' priorities concerning health research: the case of asthma and copd research in the netherlands. *Health Expectations*, 8(3):253–263, 2005.

Edward Fry. Readability versus leveling. *The reading teacher*, 56(3):286–291, 2002.

Preetinder S Gill, Tejkaran S Gill, Ashwini Kamath, and Billy Whisnant. Readability assessment of concussion and traumatic brain injury publications by centers for disease control and prevention. *International journal of general medicine*, 5:923, 2012.

K. Gillies, W. Huang, Z. Skea, J. Brehaut, and S. Cotton. 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, 2014.

Samuel D Gosling and Winter Mason. Internet research in psychology. *Annual review of psychology*, 66:877–902, 2015.

Mark Graham, Isis Hjorth, and Vili Lehdonvirta. Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research*, 23(2):135–162, 2017a.

Mark Graham, Vili Lehdonvirta, Alex Wood, Helena Barnard, Isis Hjorth, and Peter D Simon. The risks and rewards of online gig work at the global margins. 2017b.

Philipp Grandeit, Carolyn Haberkern, Maximiliane Lang, Jens Albrecht, and Robert Lehmann. Using bert for qualitative content analysis in psychosocial online counseling. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 11–23, 2020.

William Scott Gray and Bernice Elizabeth Leary. What makes a book readable. 1935.

S. Grimes. A brief history of text analytics. *BeyeNetwork, October*, 30:2007, 2007.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

Robert Gunning et al. Technique of clear writing. 1952.

Bente Hamnes, Yvonne van Eijk-Hustings, and Jette Primdahl. Readability of patient information and consent documents in rheumatological studies. *BMC medical ethics*, 17(1):1–9, 2016.

Mats O Hansson. Balancing the quality of consent. *Journal of Medical Ethics*, 24(3): 182–187, 1998.

Theodore L Harris and Richard E Hodges. *The literacy dictionary: The vocabulary of reading and writing.* ERIC, 1995.

P. Healy, S. Galvin, P.R. Williamson, S. Treweek, C. Whiting, B. Maeso, et al. Identifying trial recruitment uncertainties using a james lind alliance priority setting partnership–the priority (prioritising recruitment in randomised trials) study. *Trials*, 19:147, 2018.

Amy S Hedman. Using the smog formula to revise a health-related document. *American Journal of Health Education*, 39(1):61–64, 2008.

Ian Hodder. The interpretation of documents and material culture. *Sage biographical research*, 1, 1994.

Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

HRA. Consent & participant information sheet preparation guidance, 2014-03-03.

HRA. Appliying a proportionate approach to the process of seeking consent - hra guidance. Technical report, The University of Manchester. HRA, 2017-01-17.

HRA. Is my study research?, 2020.

Steven A Julious. Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 4(4):287–291, 2005.

Georgios A Karamitros, Nikolaos A Kitsos, and Stamatis Sapountzis. Systematic review of quality of patient information on phalloplasty in the internet. *Aesthetic plastic surgery*, 41(6):1426–1434, 2017.

J Peter Kincaid, James A Aagard, John W O'Hara, and Larry K Cottrell. Computer readability editing system. *IEEE Transactions on Professional Communication*, (1): 38–42, 1981.

J Peter Kincaid, Richard Braby, and John E Mears. Electronic authoring and delivery of technical information. *Journal of instructional development*, 11(2):8–13, 1988.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

Frank Kleemann, G Günter Voß, and Kerstin Rieder. Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science, technology & innovation studies*, 4(1):5–26, 2008.

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

Peter Knapp, David K Raynor, Jonathan Silcock, and Brian Parkinson. Can user testing of a clinical trial patient information sheet make it fit-for-purpose?-a randomized controlled trial. *BMC medicine*, 9(1):1–12, 2011b.

Siegfried Kracauer. The challenge of qualitative content analysis. *Public opinion quarterly*, pages 631–642, 1952.

Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433, 2004.

Kottur SG Arul Kumaran, Sivanandy Palanisamy, and Aiyalu Rajasekaran. Development and implementation of patient information leaflets in diabetes mellitusjphs_6 85.. 89. *Journal of Pharmaceutical Health Services Research*, 1:85–89, 2010.

Matthew LeBrun, Jason DiMuzio, Brittany Beauchamp, Susanne Reid, and Vicky Hogan. Evaluating the health literacy burden of canadaâ€™s public advisories: a comparative effectiveness study on clarity and readability. *Drug safety*, 36(12):1179–1187, 2013.

Philip Ley and Tony Florio. The use of readability formulas in health care. *Psychology, Health & Medicine*, 1(1):7–28, 1996.

Fang Liu, Sarah Abdul-Hussain, Shams Mahboob, Vijay Rai, and Andrzej Kostrzewski. How useful are medication patient information leaflets to older adults? a content, readability and layout analysis. *International journal of clinical pharmacy*, 36(4): 827–834, 2014.

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

Hans Peter Luhn. A business intelligence system. *IBM Journal of research and development*, 2(4):314–319, 1958.

Sheila MacDonald, TM McMillan, and Jacqueline Kerr. Readability of information leaflets given to attenders at hospital with a head injury. *Emergency Medicine Journal*, 27(4):279–282, 2010.

Hello Magazine. Circulation and readership, 2018.

G. Martin and J.J. Pear. Behavior modification: What it is and how to do it, 2015.

Graham P Martin. Representativeness, legitimacy and power in public involvement in health-service management. *Social science & medicine*, 67(11):1757–1765, 2008.

Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12 (8):639–646, 1969.

Glenda M McClure. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, (1):12–15, 1987.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980. ISSN 00359246.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

MHRA. Best practice guidance on patient information leaflets. 2016.

S.M. Mohammad and P.D. Turney. Nrc emotion lexicon. tech. rep., nrc technical report, 2013.

L. Moore and J. Savage. Participant observation, informed consent and ethical approval. *Nurse Researcher*, 9:58–69, 2002.

Leslie M Moore. The basic practice of statistics, 1996.

B. Moult, L.S. Franck, and H. Brady. 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, 2004.

MRC. Consent and participant information sheet preparation guidance, 2016.

Mary E Mumford. A descriptive study of the readability of patient information leaflets designed by nurses. *Journal of Advanced Nursing*, 26(5):985–991, 1997.

Emad Eldin Munsour, Ahmed Awaisu, Mohamed Azmi Ahmad Hassali, Sara Darwish, and Einas Abdoun. Readability and comprehensibility of patient information leaflets for antidiabetic medications in qatar. *Journal of Pharmacy Technology*, 33(4):128–136, 2017.

Peggy W Murphy and Terry Connally Davis. When low literacy blocks compliance. *Rn*, 60(10):58–63, 1997.

Kimberly A Neuendorf and Anup Kumar. Content analysis. *The international encyclopedia of political communication*, pages 1–10, 2015.

NHS. Clinical trials, 2017a.

NHS. Patient and public involvement payment, 2017b.

NHS. School for primary care research. 2017c.

S. Nicholls, M. Hankins, C. Hooley, and H. Smith. 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, 2009.

Lydia O'Sullivan, Prasanth Sukumar, Rachel Crowley, Eilish McAuliffe, and Peter Doran. Readability and understandability of clinical research patient information leaflets and consent forms in ireland and the uk: a retrospective quantitative analysis. *BMJ open*, 10(9):e037994, 2020.

Kevin M O'Neil, Steven D Penrod, and Brian H Bornstein. Web-based research: Methodological variables' effects on dropout and sample characteristics. *Behavior Research Methods, Instruments, & Computers*, 35(2):217–226, 2003.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

UK Parliament. The medicines for human use (clinical trials) regulations, 2004.

Silvia Pfeiffer, Stephan Fischer, and Wolfgang Effelsberg. Automatic audio content analysis. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 21–30, 1997.

Thomas Ploug and Soren Holm. Informed consent and routinisation. *Journal of Medical Ethics*, 39(4):214–218, 2013.

R. Plutchik. Emotions: A general psychoevolutionary theory. approaches to emotion, 1984, 197-219, 1984.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. Alberto: Modeling italian social media language with bert. *IJCoL. Italian Journal of Computational Linguistics*, 5(5-2):11–31, 2019.

T. Poplas-Susíc, Z. Klemenc-Ketis, and J. Kersnik. Usefulness of the patient information leaflet (pil) and information on medicines from professionals: a patients' view. a qualitative study. *Zdravni ski Vestnik*, 83:368–375, 2014.

MB Pringle, BG Natesh, and KM Konieczny. Patient information leaflet on mastoid surgery risks: assessment of readability and patient understanding. *The Journal of laryngology and otology*, 127(11):1078, 2013.

Y. Qu, J. Shanahan, and J. Wiebe. Exploring attitude and affect in text: Theories and applications. aaai spring symposium) technical report ss-04-07. aaai press, menlo park, ca, 2004.

J. Raftery, A. Young, L. Stanton, R. Milne, A. Cook, D. Turner, and P. Davidson. Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the national institute for health research health technology assessment programme, 2015.

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

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

Daniel Riffe, Charles F Aust, and Stephen R Lacy. The effectiveness of random, consecutive day and constructed week sampling in newspaper content analysis. *Journalism quarterly*, 70(1):133–139, 1993.

Ian Roberts, David Prieto-Merino, Haleema Shakur, Iain Chalmers, and Jon Nicholl. Effect of consent rituals on mortality in emergency care research. *The Lancet*, 377 (9771):1071–1072, 2011.

Elaine Robinson and David McMenemy. â€˜to be understood as to understandâ€™: a readability analysis of public library acceptable use policies. *Journal of Librarianship and Information Science*, 52(3):713–725, 2020.

Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. 2010.

Steven G Rothrock, Ava N Rothrock, Sarah B Swetland, Maria Pagane, Shira A Isaak, Jake Romney, Valeria Chavez, and Silvio H Chavez. Quality, trustworthiness, readability, and accuracy of medical information regarding common pediatric emergency medicine-related complaints on the web. *The Journal of emergency medicine*, 57(4): 469–477, 2019.

Raymol Thomas Roy, M Sonal Sekhar, Gabriel Sunil Rodrigues, and V Rajesh. Preparation and readability assessment of patient information leaflets for diabetic foot ulcers. *Journal of Social Health and Diabetes*, 1(02):079–081, 2013.

J. Saldaña. The coding manual for qualitative researchers, 2015.

Kari Sand-Jecklin. The impact of medical terminology on readability of patient education materials. *Journal of community health nursing*, 24(2):119–129, 2007.

Fernando Santos. The role of emotion in clinical text, 2017.

W.S. Sarle. Algorithms for clustering data. *Technometrics*, 32:227–229, 1990.

Eric Schenk, Claude Guittard, et al. Crowdsourcing: What can be outsourced to the crowd, and why. In *Workshop on open source innovation, Strasbourg, France*, volume 72, page 3. Citeseer, 2009.

John Seely. *Oxford Guide to Effective Writing and Speaking: How to Communicate Clearly*. OUP Oxford, 2013.

Sonal Sekhar, MK Unnikrishnan, Navya Vyas, and Gabriel Sunil Rodrigues. Development and evaluation of patient information leaflet for diabetic foot ulcer patients. *International journal of endocrinology and metabolism*, 15(3), 2017.

Danielle N Shapiro, Jesse Chandler, and Pam A Mueller. Using mechanical turk to study clinical populations. *Clinical psychological science*, 1(2):213–220, 2013.

Lucius Adelno Sherman. *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn, 1893.

Julius Sim and Martyn Lewis. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *Journal of clinical epidemiology*, 65(3):301–308, 2012.

Edgar A Smith and J Peter Kincaid. Derivation and validation of the automated readability index for use with technical materials. *Human factors*, 12(5):457–564, 1970.

Statista. Monthly reach of daily mail and the mail on sunday newspapers in great britain from april 2018 to march 2019, by demographic group, 2018.

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

A Suhaj, Aswini Kumar Mohapatra, Rahul Magazine, MK Unnikrishnan, V Rajesh, Sonal Sekhar, et al. Development and readability assessment of patient information leaflets for chronic obstructive pulmonary disease. *Indian Journal of Pharmaceutical Education and Research*, 5(2), 2015.

Randall S Sumpter. News about news: John g. speed and the first newspaper content analysis. *Journalism History*, 27(2):64–72, 2001.

Erik Nathan Swartz. The readability of paediatric patient information materials: Are families satisfied with our handouts and brochures? *Paediatrics & child health*, 15(8): 509–513, 2010.

Giuseppina Terranova, Marcello Ferro, Clara Carpeggiani, Virginia Recchia, Larissa Braga, Richard C Semelka, and Eugenio Picano. Low quality and lack of clarity of current informed consent forms in cardiology: how to improve them. *JACC: Cardiovascular Imaging*, 5(6):649–655, 2012.

Edward L Thorndike. The teacher's word book. 1921.

Jeffrey S Tobias and Robert L Souhami. Fully informed consent can be needlessly cruel. *BMJ: British Medical Journal*, 307(6913):1199, 1993.

Connie K Varnhagen, Matthew Gushta, Jason Daniels, Tara C Peters, Neil Parmar, Danielle Law, Rachel Hirsch, Bonnie Sadler Takach, and Tom Johnson. How informed is online informed consent? *Ethics & Behavior*, 15(1):37–48, 2005.

Y. Volcani and D. Fogel. System and method for determining and controlling the impact of text, 2006.

Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583, 2018.

Robert Philip Weber. *Basic content analysis*. Number 49. Sage, 1990.

Marilyn Domas White and Emily E Marsh. Content analysis: A flexible methodology. *Library trends*, 55(1):22–45, 2006.

ICK Wong. Readability of patient information leaflets on antiepileptic drugs in the uk. *Seizure*, 8(1):35–37, 1999.

H. Xiong, J. Wu, and J. Chen. K-means clustering versus validation measures: a data-distribution perspective. ieee transactions on systems, man, and cybernetics, part b (cybernetics), 39, 318-331, 2009.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE international conference on data mining*, pages 427–434. IEEE, 2003.

Omar Zaidan and Chris Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1220–1229, 2011.