United States Food and Drug Administration

Text Analysis of Proprietary Drug Name Interpretations

OMB Control No. 0910-NEW

SUPPORTING STATEMENT

Part B. Statistical Methods

1. Respondent Universe and Sampling Methods

Pretest and main study general population participants will be recruited from Ipsos Public Affairs' KnowledgePanel®. Ipsos provides a nationally representative sample recruited via address-based sampling (ABS), as opposed to online opt-in panels that recruit participants via online advertising. The availability of pre-collected profile data, which include standard demographic and health data for members, allows Ipsos to target respondents of interest and obtain a diverse set of survey completes.

For the primary care physician (PCP) main study sample, participants will be recruited using a two-stage approach that will begin with a purchased list of PCPs based on the AMA Physician Masterfile. These members will then be matched to one or more sample provider lists to recruit participants for this study. This process will be used for the main study data collection only. The pretest sample will be obtained from Ipsos' partner OptIn sample sources because of the small sample size and different objectives of the pretests, which do not need to be generalizable.

The PCP recruitment plan must balance a few challenges. Because PCPs are not nearly as numerous as consumers in the general population, relying on probabilistic recruitment methods such as random-digit-dial or ABS would likely not be feasible from a cost, burden, and timeline perspective. Thus, online physician panels represent the best way to target and obtain survey completes from PCPs. At the same time, relying on a single physician panel might not be conducive to making generalizable statements about PCPs given biases that exist within any single panel. Our recruitment method will address these biases by starting with a sample that is representative of the AMA Masterfile's PCP composition. Ipsos will then work with various online physician panels to match their member base against the sampled list using each record's unique identifier.

Participants from pretesting will not be allowed to participate in the main study. The sample will be limited to English-only, with a range of representation in terms of age, gender, race, ethnicity, and education. Additional eligibility criteria include having not participated in a focus group or interview in the past 3 months, not working in marketing or employed by the U.S. Department of Health and Human Services, not working in health care (for consumers only), and engaging in patient care at least 50% of the time (for PCPs only).

2. Procedures for Collection of Information

Our sample will include 300 general population consumers and 300 PCPs. We have designed a within-subjects experiment in which participants will be exposed to multiple drug names to maximize power to find differences with this sample size. The stimuli will comprise 60 experimental names and 60 control names. Participants will be randomized to 1 of 10 groups so that no one responds to more than 12 names in total. Each participant will see six experimental names and six control names. The experimental names will be names with suspected promotional implications, whereas the control names will not have suspected promotional implications. Names will be viewed in random order. Participants will respond in open-ended text boxes about their perceptions of each drug name. Supplementary closed-ended questions may also be presented. We will conduct text analysis of the responses and present descriptive results for individual drug names by participant cohort (i.e., consumers versus PCPs), and we will also code and compare responses across types of drug names.

Power Analysis

The proposed study will use a total of 120 stimuli to represent two broad classes of proprietary names that potentially imply (to varying degrees) promotional implications that extend beyond the indication: (1) control names ($k = 60$) made up of random combinations of syllables; and (2) names with promotional implications ($k = 60$).

Because it would be impractical to ask participants to respond to all 120 names, we will partition the full set of names into 10 mutually exclusive blocks of 12 names (6 control; 6 experimental) and randomly assigning participants so that everyone interprets the same list of names in their respective blocks. Over the whole study, we will have collected data on the full set of names by adopting this stimuli-within-block design (Westfall, Kenny, & Judd, 2014), which greatly improves statistical power without overburdening individual participants. The analysis will focus on potential differences in interpretation that arise categorically between control names and those with promotional implications, rather than between any specific pairs of names. Unlike more conventional study designs where participants are the only random factor, the proposed study design treats stimuli as a second random factor, which distributes error across the stimuli and improves confidence that the findings result from the categorical distinction of interest rather than the specific messages or stimuli used in a study (Judd, Westfall, & Kenny, 2016). Thus, an added benefit of including more drug names and treating them as a random factor is that it protects against the "language-as-fixed effects fallacy" (Clark, 1973), or the "case-category confound" (Jackson, 1992), which is a threat to generalization that arises because particular stimuli may vary on dimensions other than the characteristics shared by members in the categories of interest, leaving an alternative explanation for observed effects.

Given an alpha level of 0.05, a sample size of 600 participants who each interpret 12 names drawn from a total of 120 proprietary names (60 control, 60 with promotional implications) we will have 90% power to detect small-to-medium effects by experimental condition ($d \geq 0.29$), participant cohort ($d \geq 0.23$), and the interaction of experimental condition by

participant cohort ($d \geq 0.22$). Importantly, increasing the number of participants in a design with two random factors has less of an impact on power to detect differences by experimental condition than increasing the number of stimuli and, all else being equal, power reaches asymptote at fairly low participant sample sizes (Judd, Westfall, & Kenny, 2016). For example, we ran power analysis for up to 24,300 participants and found that increasing the number of participants beyond 300 resulted in modest improvements in power for this study. Treating the consumer and PCP samples separately, respective sample of 300 participants will have 90% power to detect a medium-small effect by experimental condition ($d \geq 0.32$).

Weighting

Ipsos plans to design weights for the two samples to account for any under- and overrepresentation of demographic groups in the final survey completes. For this project, the PCP weighting targets will be derived from the AMA Masterfile. This approach was used successfully on the Healthcare Provider Survey of Prescription Drug Promotion (OMB Control No. 0910-0730). The consumer weights will be developed based on the Current Population Survey, which are developed using data from the U.S. Census.

Analysis

Topic Modeling with Treatment Variables and Sentiment Analysis

Topic modeling (Blei, 2012) is an unsupervised text method that produces categories directly from text, usually using a Bayesian generative probabilistic model to jointly estimate the composition of words within topics and the proportion of topics within documents. We will examine and present descriptive results for individual names. However, given our goals of understanding promotional implications of prescription drug names across consumers and PCPs, we are also interested in whether there are differences in topic distributions across our treatment and control arms (control vs. promotional implications) and between populations (consumers and PCPs).

To achieve this, we will use a two-step approach. In the first step, we will use topic modeling to measure how often a topic is discussed ("prevalence") and the language used to discuss the topic ("content") per document (i.e., participants' open-ended responses to each name). The output of this procedure will yield topic variables that record the proportion of a document discussing each topic. Additionally, the models also provide information about the frequency of words associated with each topic, which will allow our research team to assign each topic a descriptive name encoding semantic meaning. The second step will use the topic proportions (now assigned to each open-ended response) to compare differences in topic proportions between responses for our treatment and control names and by participant cohort.

Validating Topic Models. There are several common estimation and interpretational challenges with traditional topic models. In particular, these include the need to determine an appropriate number of topics, to find stable topics across different initialization

schemes, and to correct for perceived poor quality or missing topics (Chang et al., 2009; Lee et al., 2017; Mimno et al., 2011).

To address these challenges, we propose several validation steps, both formal and informal, including assessments of face validity and goodness-of-fit comparisons. Specifically, we will calculate coherence metrics to assess model fit, as well as perform validation exercises to assess if the generated topics can be easily interpreted. The results from these validation steps will be included in dissemination of the study results to aid transparency.

If after running the validation steps we still consider the topic coherence and specificity inadequate, we can explore semi-supervised topic models, such as CorEx (Gallagher et al., 2017). This topic model allows the research team to inject domain knowledge via "anchor words" to guide the model toward otherwise underrepresented but important topics.

Unsupervised Learning with Short Text Responses. Although topic modeling is an ideal method under the right conditions (see Banks et al., 2018 and Roberts et al., 2014), our prior experience suggests that it is not always possible to develop a topic model that can detect nuanced concepts relevant to a targeted research question. This concern is especially acute when working with a small number of documents or documents with a limited amount of text (which is a potential concern on this study if participants do not enter much in the text boxes).

If we discover in pre-testing that respondents do not provide lengthy reactions to drug names, we can explore unsupervised clustering methods tailored to short text responses.

Validating clustering output is similar in many ways to topic modeling and they share many researcher design choices, such as choosing the number of topics/clusters. To help simplify these choices, we propose using a density-based clustering method (McInnes et al., 2017) that selects the number of clusters empirically from the data. With density-based clustering methods, the choice of the number of clusters is swapped for parameters describing how dense local points should be to be considered a cluster.

If we still have difficulty identifying stable topics using unsupervised methods, we propose a supplementary approach called weak supervision.

Weak Supervision. Unsupervised methods, like topic modeling and clustering, can be useful when there are not known categories of interest, but research teams often have specific hypotheses or guidance to help determine categories of interest. Supervised learning methods are designed to classify new observations into known categories of interest. To do this, supervised learning models use examples of coded documents to learn how to code new documents. Supervised methods have two main advantages over unsupervised methods: (1) they are optimized specifically to distinguish between categories of interest; and (2) they are much easier to validate, with clear statistics that summarize model performance (e.g., accuracy, precision, recall).

Normally, the data to develop a supervised classifier would come from human-annotated

documents, similar to what would be created during a thematic content analysis. To create a training data set that mimics this required input without a formal coding exercise, we propose a weak supervision approach called data programming (Ratner et al., 2016; 2020) to create codes for sentences thought to contain the topics of interest.

Sentiment Analysis. As a final analysis, we will detect underlying sentiment for individual names elicited from respondents. We will run several sentiment analysis models to determine the best fit for our use.

Comparing Topics and Sentiment by Treatment and Control Arms and by Participant Cohort. We will conduct multilevel mixed-effects regression analyses to understand the effect of proprietary drug names with promotional implications on participants' interpretations of drug efficacy and risk (i.e., topic prevalence and sentiment measures derived from topic modeling and sentiment analysis models). Our statistical approach will adjust for data-correlated errors that are likely to arise because of multiple measurements made on the same subject. Hierarchal or mixed models will be used to account for variability among and within participants from measure to measure.

Specifically, the proposed stimuli-within-blocks study design can be thought of as a multilevel model with nested random effects: stimuli (i.e., proprietary drug names) are nested within experimental condition (i.e., control names vs. names with promotional implications) and participants. Experimental condition is a Level-1 fixed factor, and both participants and stimuli are Level-2 random factors. A general example of the analytic model for this design is described by Judd, Westfall, & Kenny (2016). The comparative effects of specific names used as stimuli are not the inferential focus of this statistical analysis; instead the names have been chosen for the sake of ecological validity to represent a variety of control drug names and proprietary drug names that have or could conceivably appear on the market.

In these models, we will estimate a topic prevalence or sentiment outcome (e.g., overstatement of efficacy) using experimental condition as the main predictor of interest.

3. Methods to Maximize Response Rates and Deal with Non-response

Strategies to maximize response rates for the data collection include:

- Repeated follow-ups
- Use of a panel of respondents accustomed to participating in periodic surveys
- Administer survey over the Internet, for convenience
- Survey length of 20 minutes or less
- Mobile optimization for web surveys

Obtaining an adequate response rate is a well-known challenge in survey research in recent years. Much research has been devoted to finding ways to maximize response rates to ensure high quality data (Czajka & Beyler, 2016). The strategies we propose are based on the most recent literature on enhancing response.

Similarly, obtaining an adequate response from a sample of physicians is a well-documented challenge. Physicians are frequently asked to take part in surveys about medical issues and, given their busy schedules, are often more reluctant than other types of respondents to participate (Kablunde et al., 2012). We will draw from our experience with various strategies to maximize response rates as summarized below:

- Prepaid incentive checks
- Large incentives (see section 9 of Supporting Statement A)
- Use of an existing panel of participants accustomed to participating in periodic surveys
- Administer the survey over the Internet, for convenience
- Survey length of 20 minutes or less

4. Test of Procedures or Methods to be Undertaken

Data collection will begin with two small pretests in order to test and evaluate procedures. The pretests will utilize the same procedures and methods as the full data collection but utilize only a small sample.

5. Individuals Consulted on Statistical Aspects and Individuals Collecting and/or Analyzing Data

The contractor, RTI International, will collect and analyze data on behalf of FDA as a task order under Contract 75F40120A00017. Bridget J. Kelly, Ph.D., MPH, is the Project Director, (202) 728-2098. Review of contractor deliverables and supplemental analyses will be provided by the Research Team, Office of Prescription Drug Promotion (OPDP), Office of Medical Policy, CDER, FDA, and coordinated by Kevin R. Betts, Ph.D., (240) 402-5090, and Amie O'Donoghue, Ph.D., (301) 796-1200.

References

Banks, G.C., Woznyj, H.M., Wesslen, R.S., and Ross, R.L. "A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App). *Journal of Business and Psychology*, 33(4):445–459, 2018.

Blei, D.M. "Probabilistic Topic Models." *Communications of the ACM*, 55(4), 77–84, 2012.
Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., … Kurzweil, R. (2018). Universal Sentence Encoder. ArXiv:1803.11175 [Cs]. http://arxiv.org/abs/1803.11175

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., and Blei, D.M. "Reading Tea Leaves: How Humans Interpret topic models." In Y. Bengio et al. (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288–296). Curran Associates, Inc, 2009.

Czajka, J.L. and A. Beyler. "Background Paper: Declining Response Rates in Federal Surveys: Trends and Implications." Mathematica Policy Research, 2016. Available at: https://aspe.hhs.gov/system/files/pdf/255531/Decliningresponserates.pdf

Clark, H.H. "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359, 1973.

Gallagher, R.J., Reing, K., Kale, D., and Steeg, G. V. "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." *Transactions of the Association for Computational Linguistics*, *5*(0), 529–542, 2017.

Jackson, S. *Message Effects Research: Principles of Design and Analysis*. New York, NY: Guilford, 1992.

Judd, C.M., Westfall, J., and Kenny, D.A. "Experiments with More than One Random Factor: Designs, Analytic models, and Statistical Power." *Annual Review of Psychology*, 68, 17.1–17.25, 2016.

Kablunde, C., Willis, G.B., McLeod, C.C., Dillman, D.D., Johnson, T.P., Greene, S.M., and M.L. Brown. "Improving the Quality of Surveys of Physicians and Medical Groups: A Research Agenda." *Evaluation and the Health Professions*, *35*(4), 477–506, 2012.

Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., and Findlater, L. "The Human Touch: How Non-Expert Users Perceive, Interpret, and Fix Topic Models." *International Journal of Human-Computer Studies*, 105, 28–42, 2017.

McInnes, L., Healy, J. and Astels, S. "Hierarchical density-based clustering." *Journal of Open Source Software, The Open Journal*, 2(1), 2017.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272), 2011.

Ratner, A. J., De Sa, C.M., Wu, S., Selsam, D., and Ré, C. "Data Programming: Creating Large Training Sets, Quickly." In D. D. Lee et al. (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3567–3575), 2016.

Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel. "Rapid Training Data Creation with Weak Supervision." *The VLDB Journal*, 29(2), 709–730, 2020.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., & Rand, D. G. Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science*, 58(4), 1064–1082, 2014.

Westfall, J., Kenny, D.A., & Judd, C.M. "Statistical Power and Optimal Design in Experiments in Which Samples of Participants Respond to Samples of Stimuli." *Journal of Experimental Psychology: General,* 143(5), 2020–2045, 2014.