Smarter micro-targeting to improve global health outcomes: scaling cluster segmentation on novel types of data for precision public health

Elisabeth Engl
Surgo Foundation
641 S Str NW
Washington, DC 20001
elisabethengl@surgofoundation.org

Sema K. Sgaier
Surgo Foundation
641 S Str NW
Washington, DC 20001
semasgaier@surgofoundation.org

Abstract

In global health, optimized interventions that rely on voluntary uptake should account for decision-making heterogeneity within the target population. This is a major challenge: to efficiently use scarce resources, the right people should be targeted with the right intervention at the right time. This entails segmenting the population of interest based on the differences that underlie their decision-making, prioritizing which segments to target, and then developing segment-specific intervention strategies. Effective global health interventions at scale require large financial and human resource investments, and both are limited in resource-poor settings. Therefore, scaling segmentation approaches to optimize intervention development and deployment is a key challenge and requires collaboration at the intersection of global health and machine learning experts. Here, we demonstrate how the collection of novel types of data coupled with the application of unsupervised classification algorithms can help micro-target interventions on the ground to drive health outcomes. Highlighting a case study on Voluntary Medical Male Circumcision in Africa, we discuss lessons learned and innovations in future programs, especially how mobile approaches to data collection and analytics can be streamlined to make segmentation a scalable approach across global health.

1 Introduction

Targeted interventions are a key component of precision public health. Using one-size-fits-all approaches encouraging the uptake of Voluntary Medical Male Circumcision (VMMC), a key component of national-scale HIV prevention programs recommended by the WHO and scaled up in 14 countries, uptake of the procedure initially increased, but soon plateaued. To close that gap, previous work by Sgaier et al. (2017) segmented men on their drivers and barriers to VMMC to understand their decision-making, and national governments in Zambia and Zimbabwe are now piloting different interventions to encourage men to elect the procedure (Sgaier et al. 2017). Such fault lines can be uncovered by unsupervised cluster segmentation techniques. The algorithms themselves are a staple of machine learning and have been refined over the years to suit a variety of data types and structures (Huang 1998, Broder et al. 2014, Peterson et al. 2017). However, knowledge of analytical techniques available is not enough, and application of the approach in global health programs has been almost non-existent (Sgaier et al. 2018). Cluster analysis follows the 'garbage in, garbage out' principle: the data segmented on determine the patterns that are uncovered and, ultimately, the interventions that can be designed. While customer segmentation has a long history of use both in the public sector and private market research (Yankelovich and Meer 2006), robust psycho-behavioral segmentation capturing differences not only in demographic or socio-economic

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

factors, but also variation in structural barriers (such as access to services), behavioral patterns (such as other health behaviors), beliefs around costs and benefits, and influencers (Slater et al. 2006, Sgaier et al. 2018), is an emerging field (Matz et al. 2017, Sgaier et al. 2017). Holistic data to construct such actionable segments is not only difficult to define, but also to measure. Here, we highlight a case study demonstrating how these types of data can be obtained and utilized more readily, and the challenges that remain to scale psycho-behavioral segmentation. In global health, both expanding the types of data used and applying appropriate machine learning clustering techniques are novel, and expertise is scarce: successfully developing actionable segments requires close collaboration between program domain experts, data scientists, and behavioral scientists. Based on lessons learned from the VMMC program, we then introduce new case studies innovating on faster and more scalable data collection, analysis, and deployment of clustering techniques. We segment household members with the goal of improving the rates of institutional delivery in Uttar Pradesh, and women on their drivers and barriers to family planning in Madhya Pradesh, both in India. We highlight three recommendations: 1) data collection should from the get-go be designed to be suitable for clustering algorithms, and incorporate a robust framework of categories covering a potentially wide-ranging set of decision drivers and barriers that can serve as segmentation dimensions; 2) data collection should be performed at scale, linking several actors within the system relevant to the decision; and 3) streamlining cluster algorithm selection and intervention deployment requires further technological development.

2 Data structure

The VMMC data set contained survey responses from 2,000 men in Zambia and 2,001 men in Zimbabwe. Psycho-behavioral segments were constructed based on quantitative survey data. Survey design was based on well-validated frameworks of behavioral theory structuring categories of potential contextual and perceptual drivers (Yzer 2012, Sgaier et al. 2017), as well as exploratory qualitative research (Sgaier et al. 2017). Driver categories may include infrastructure and other structural barriers, policies and laws, social norms, beliefs around costs, benefits, and risks, as well as awareness, intention, perceived influencers, and other health behaviors. Using a structured framework greatly sped up instrument development, and data scales were set up to be suitable for several unsupervised cluster algorithms.

3 Analytical Methodology

Unsupervised cluster analysis was used to construct psycho-behavioral segments. From the quantitative survey, variables with low response reliability were excluded as inputs to clustering, followed by dimensionality reduction using canonical correlation analysis. Subsequently, a set of analyses using competing and complementary algorithms (such as hierarchical, k-means and hybrid clustering) was performed iteratively. Solutions were evaluated using both statistical measures of comparing intra- and inter cluster variance and practical considerations. Finally, a CHAID-based predictive model enabled the creation of a segment allocation ('typing') tool that could be applied by field workers on the ground (for details, see Sgaier et al. 2017).

4 Results

Segmenting men according to their drivers and barriers to VMMC resulted in 6 distinct psycho-behavioral segments in Zimbabwe and 7 in Zambia (Sgaier et al. 2017). An example of key segment characteristics is given in Table 1:

Table 1: Key segment differentiators for men in Zimbabwe (modified from Sgaier et al. 2017).

Country	Segment	Key factors defining segment profiles				
Zimbabwe		Motivation/ need for VMMC	Rejection due to cognitive dissonance	Perceived lack of ability	Acceptance of social support	Personal constraints
	Enthusiasts	Strong motivation	Neutral	Average ability	Highly socially driven	Some fears
	Champions	Strong motivation	No rejection	Strong ability	Highly independent	Some fears
	Neophytes	Neutral motivation	Strong rejection	Lack of ability	Highly independent	Some fears
	Scared Rejecters	Neutral motivation	Strong rejection	Strong ability	Highly independent	Strong fears
	Embarrassed Rejecters	Weak motivation	No rejection	Average ability	Highly socially driven	Strong fears
	Highly Resistant	Weak motivation	Strong rejection	Strong ability	Highly socially driven	No fears

These segments also differed in their circumcision status and - if not circumcised - future intention, as well as in their estimated and reported risk of acquiring HIV, and could be 'profiled' or differentiated on many other behaviors or outcomes of interest. Then, segments were prioritized for intervention along criteria such as prevalence, ease of conversion, and impact on others (Sgaier et al. 2017). Furthermore, a chi-squared automatic interaction detection (CHAID) algorithm was used as a predictive tool to determine key questions that field workers could ask men while talking to them in the field (Figure 1). National-scale programs in Zimbabwe and Zambia now use such typing tools to identify the messages and decision-making aides field workers should focus on.

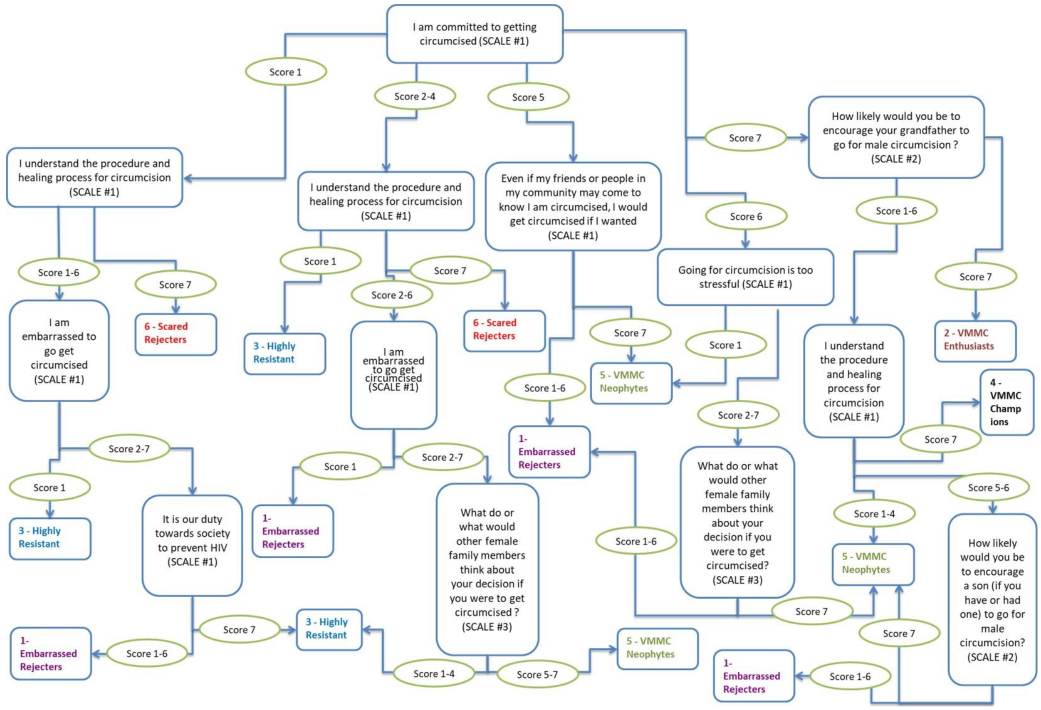
I am committed to getting circumcised (SCALE #1)
Score 1
Score 2-4
Score 5
Score 7
Score 6
I understand the procedure and healing process for circumcision (SCALE #1)
I understand the procedure and healing process for circumcision (SCALE #1)
Even if my friends or people in my community may come to know I am circumcised, I would get circumcised if I wanted (SCALE #1)
How likely would you be to encourage your grandfather to go for male circumcision ? (SCALE #2)
Score 1-6
Score 7
Score 1
Score 7
Score 2-6
Going for circumcision is too stressful (SCALE #1)
Score 1-6
Score 7
I am embarrassed to go get circumcised (SCALE #1)
6 - Scared Rejecters
3 - Highly Resistant
I am embarrassed to go get circumcised (SCALE #1)
6 - Scared Rejecters
Score 7
Score 1
2 - VMMC Enthusiasts
5 - VMMC Neophytes
I understand the procedure and healing process for circumcision (SCALE #1)
4 - VMMC Champ ions
Score 1-6
Score 2-7
1- Embarrassed Rejecters
Score 2-7
Score 7
Score 1
Score 2-7
What do or what would other female family members think about your decision if you were to get circumcised? (SCALE #3)
Score 5-6
Score 1-4
3 - Highly Resistant
It is our duty towards society to prevent HIV (SCALE #1)
1- Embarrassed Rejecters
What do or what would other female family members think about your decision if you were to get circumcised ? (SCALE #3)
Score 1-6
5 - VMMC Neophytes
How likely would you be to encourage a son (if you have or had one) to go for male circumcision? (SCALE #2)
Score 7
Score 7
3 - Highly Resistant
1- Embarrassed Rejecters
Score 1-6
Score 1-4
Score 5-7
5 - VMMC Neophytes
1- Embarrassed Rejecters
Score 1-6

Figure 1: CHAID-based typing tool deployed by field workers to allocate men to a segment in Zimbabwe. Scale 1 (7-point scale): 7 = 'Strongly agree'; 4 = 'Neither agree nor disagree'; 1 = 'Strongly disagree'. Scale 2 (7-point scale): 7 = 'Would definitely encourage'; 4 = 'Would neither encourage nor discourage'; 1 = 'Would definitely NOT encourage'. Scale 3 (7-point scale): 7 = 'They think I definitely should get circumcised'; 4 = 'They don't have any particular opinion'; 1 = 'They think I definitely should NOT get circumcised'. Reproduced from Sgaier et al. 2017.

5 Lessons learned and future work

Unsupervised cluster analysis on psycho-behavioral data as outlined here has been used for intervention design driving VMMC uptake for the first time on a national-scale level (Sgaier et al. 2017). Four key lessons stand out from this program: first, the type of data collected must expand to encompass a greater set of potential driver and barrier categories, and data collection methods should evolve from time-, resource-, and cost-intensive in-person methods. This type of data currently rarely exists in global health, and almost never at scale. Second, cluster algorithm selection should be automated to the degree possible, while under the supervision of analysts and program implementers. Third, mobile technologies should be leveraged at all points from data collection and analysis to intervention deployment, in order to iteratively test solutions as quickly as possible. And finally, global health must enhance its capacity to integrate machine learning experts into the design of segmentation programs early on.

In two further case studies in low-resource settings, we innovate on the types and methods of data collection and how solutions are deployed, so that cluster analysis can become a truly scalable tool for more targeted interventions.

5.1 Understanding institutional delivery in Uttar Pradesh, India

Data linkages across public and (wherever possible) private-sector data sets could reduce the need for self-report on many measures. Some data types, such as demographic or infrastructure data, are more readily available than others, including individual-level health knowledge data or beliefs around health risks and benefits. Frameworks of types of drivers and barriers can greatly facilitate the structuring of questionnaires (Sgaier et al. 2017), and we are now systematically applying this 'skeleton of drivers and barriers' for faster instrument design. To understand the drivers and barriers to institutional delivery in Uttar Pradesh, India, we not only collect holistic data on household members, with a focus on 6,000 women who recently gave birth, but also link that data to 1,500 community health workers that have visited each individual household, as well as contextual data such as infrastructure and healthcare facility audits. Our work has moved on from paper-based to mobile data collection tools, but in low-income geographies data collection still relies on in-person processes. As mobile phone usage spreads, data collection could move from intensive in-person to remote surveys; however, response quality issues can be substantial. We are also building a mobile platform for community health workers, who can in the future be targeted in real time based on behaviors and responses. Segment-specific messaging to encourage bi-directional communication between households and community health workers can be deployed directly through that platform.

5.2 Understanding drivers of family planning in Madhya Pradesh, India

Cluster analytics should move on from post-hoc 'more art than science' techniques and be iterated on in real time as the data is collected, enabling fast adaption of potential solutions. This will require advances in automated algorithm selection and tuning, with domain knowledge input from data scientists and program designers at key stages. Currently, we are developing a real-time segmentation allocation tool to deploy a targeted intervention at the end of a customer survey. Measuring a later behavioral output will estimate the quality of that targeting, iteratively fine-tuning the targeting algorithm.

6 References

[1] Sgaier, S. K. et al. A case study for a psychographic-behavioral segmentation approach for targeted demand generation in voluntary medical male circumcision. eLife 6 (2017).

[2] Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 22 (1998).

[3] Broder, A., Garcia-Pueyo, L., Josifovski, V., Vassilvitskii, S. & Venkatesan, S. Scalable K-Means by ranked retrieval. In WSDM'14 Proceedings of the 7th ACM international conference on Web search and data mining 233–242 (ACM Press, 2014).

[4] Peterson, A. D., Ghosh, A. P. & Maitra, R. Merging K-means with hierarchical clustering for identifying general-shaped groups. Stat 16 (2017).

[5] Sgaier, S. K., Engl, E. & Kretschmer, S. Time to scale psycho-behavioral segmentation in global development. Stanford University, and Stanford Center on Philanthropy and Civil Society: Stanford Social Innovation Review (2018).

[6] Yankelovich, D. & Meer, D. Rediscovering market segmentation. Harvard Business Review 24, 122–131 (2006).

[7] Slater, M. D., Kelly, K. J. & Thackeray, R. Segmentation on a shoestring: health audience segmentation in limited-budget and local social marketing interventions. Health Promotion Practice 7, 170–173 (2006).

[8] Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences 114, 12714–12719 (2017).

[9] Yzer, M. The integrative model of behavioral prediction as a tool for designing health messages. Health communication message design: Theory and practice 21–40 (2012).