PCR3

Assessing Concept Saturation and Sample Representativeness When Using Social Media Data to Inform Concept Elicitation Studies

Conrad Bessant[1,2,3], Yasemin Bridges[1], Marzana Chowdhury[1], Asiyya Tahsin[1], Maryam Abdollahyan[1,2], Fabrizio Smeraldi[1,2,3], Bill Byrom[4]

1: Queen Mary University London, London, UK. 2: Mebomine Ltd., London, UK. 3: Alan Turing Institute, London, UK. 4: Signant Health, London, UK.

INTRODUCTION

Social media (SM) data, such as that posted on online health boards (OHBs), can provide insights to aid concept elicitation in the development of clinical outcome assessments, and other application areas.

While SM cohorts can be much larger than those achieved in cognitive interview studies, understanding concept saturation remains important.

In addition, OHB users are often younger and include greater proportions of females compared to many patient populations, making it important to consider the representativeness of the sample when drawing conclusions.

METHODS

We used 383 OHB posts from 271 individuals with congestive heart failure (CHF) that were used in an AI-assisted concept elicitation exercise to understand meaningful aspects of health related to physical activity in this patient population [1].

Sample representativeness

We implemented weighting adjustments to adjust for the differences in age distributions between our cohort and the CHF population for the 17 high-level concepts identified. Differences observed could be used to help make inferences about the representativeness of findings from our sample.

Concept saturation assessment

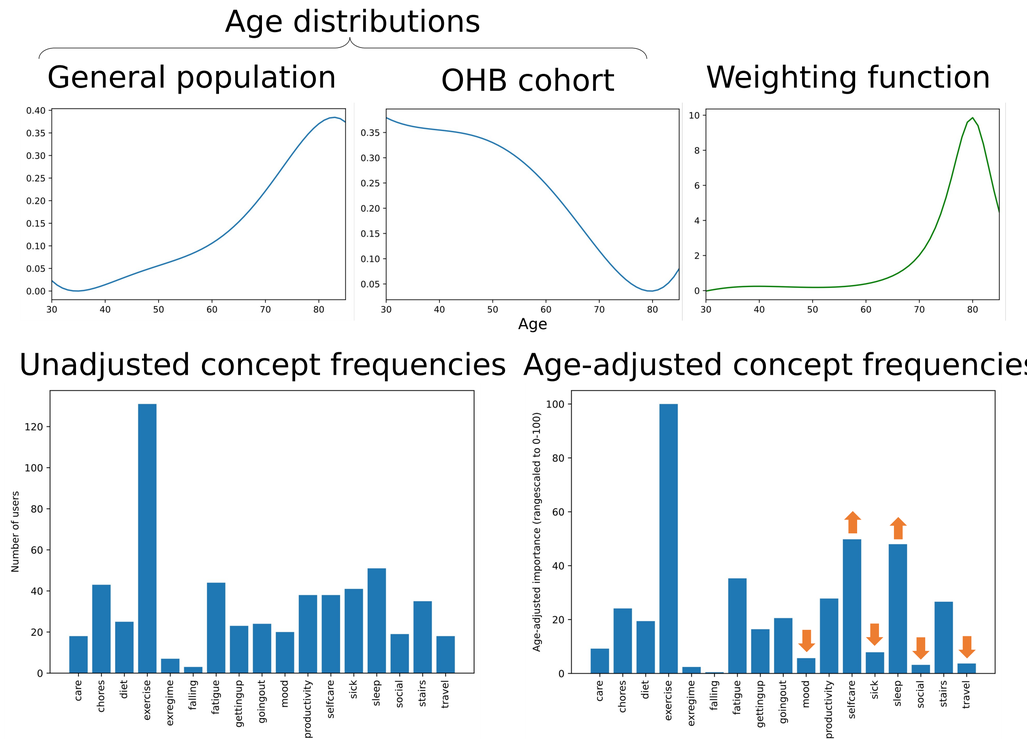
We used autocorrelation over time between sub-concepts identified within the exercise high-level concept to measure the convergence of findings to quantify concept saturation. Simulation confirmed findings independent of the order of posts in the coding process.

AGE ADJUSTMENT FOR HIGH-LEVEL CONCEPTS

Ages were determined for 209/271 (77%) of individuals, were normally distributed (range: 20 – 96, mean: 54, SD: 15 years), and lower ages were more represented than in the true population.

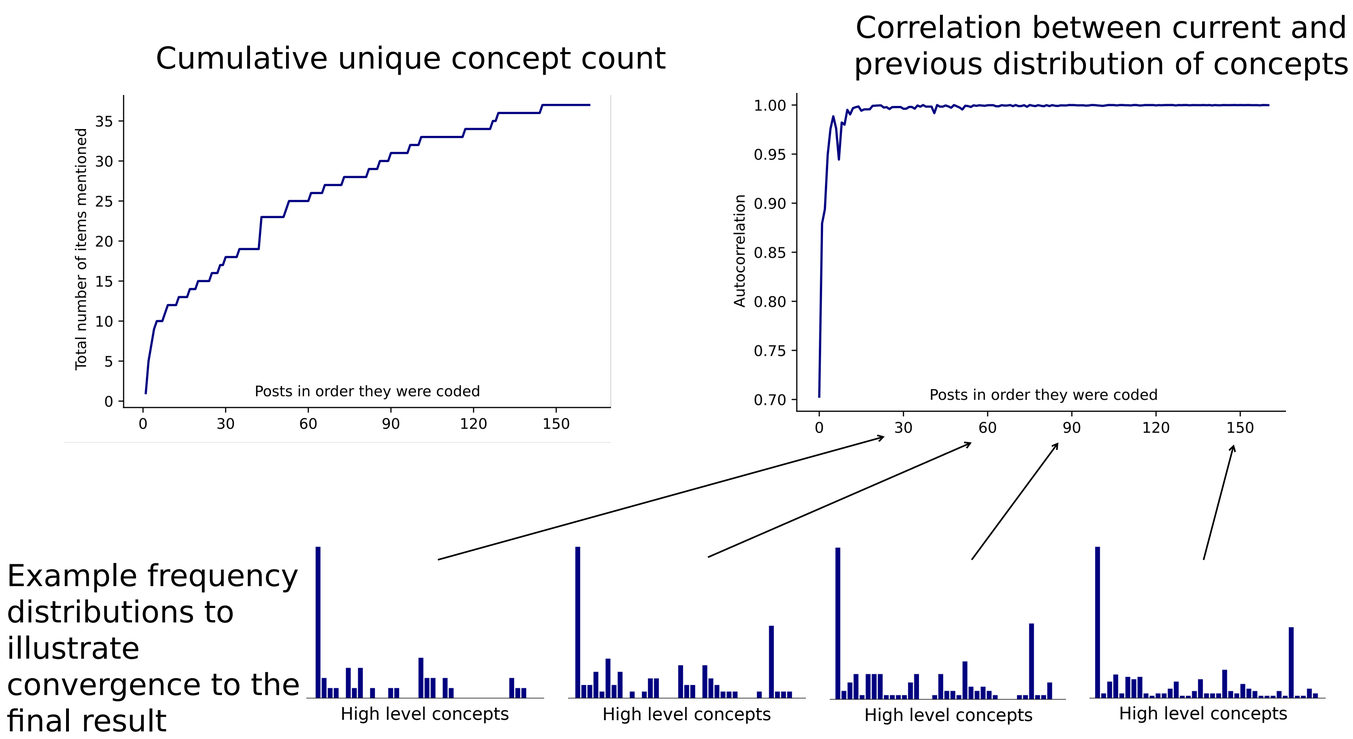
AGE ADJUSTMENT METHODOLOGY

1. We fitted a polynomial to the age distribution of OHB users as sampled, and to published population prevalence data [2].
2. We applied the ratio of these curves to weight the frequency of each concept by age.
3. Age-adjusted results were scaled to provide a relative frequency, where a concept would have a value of 100% if it was mentioned by all patients for which age could be determined.



EXERCISE-RELATED SUB-CONCEPT SATURATION ASSESSMENT

Saturation analysis showed asymptotic growth in identified concepts during the coding process, but correlation between the overall frequency distribution of concepts from post to post began to stabilise after the first 30 posts, and continuing to code beyond the first 90 posts had negligible impact on the overall findings of the study. While the results shown are for the sequence in which posts were coded in this study, analysis of 100 different permutations of this sequence demonstrated the general trends are independent of the order in which posts were coded.



Conclusions: OHB data may contribute to concept elicitation knowledge important in outcome measure development. The methods we describe help to ensure that findings are representative of the patient population, and that concept saturation is established. These approaches contribute to enhancing the robustness of findings from social listening studies.

[1] Bridges Y, Chowdhury M, Tahsin A et al. Identification of meaningful aspects of physical activity: concept elicitation by AI-assisted coding of online patient conversations . Value in Health 2022; 25: S226.
[2] Bosch L, Assmann P, de Grauw WJC et al. Heart failure in primary care: prevalence related to age and comorbidity. Prim Health Care Res Dev. 2019; 20: e79.

Presented at ISPOR Europe 2022, 06-Nov to 09-Nov 2022

Contacts: conrad@mebomine.com, bill.byrom@signanthealth.com