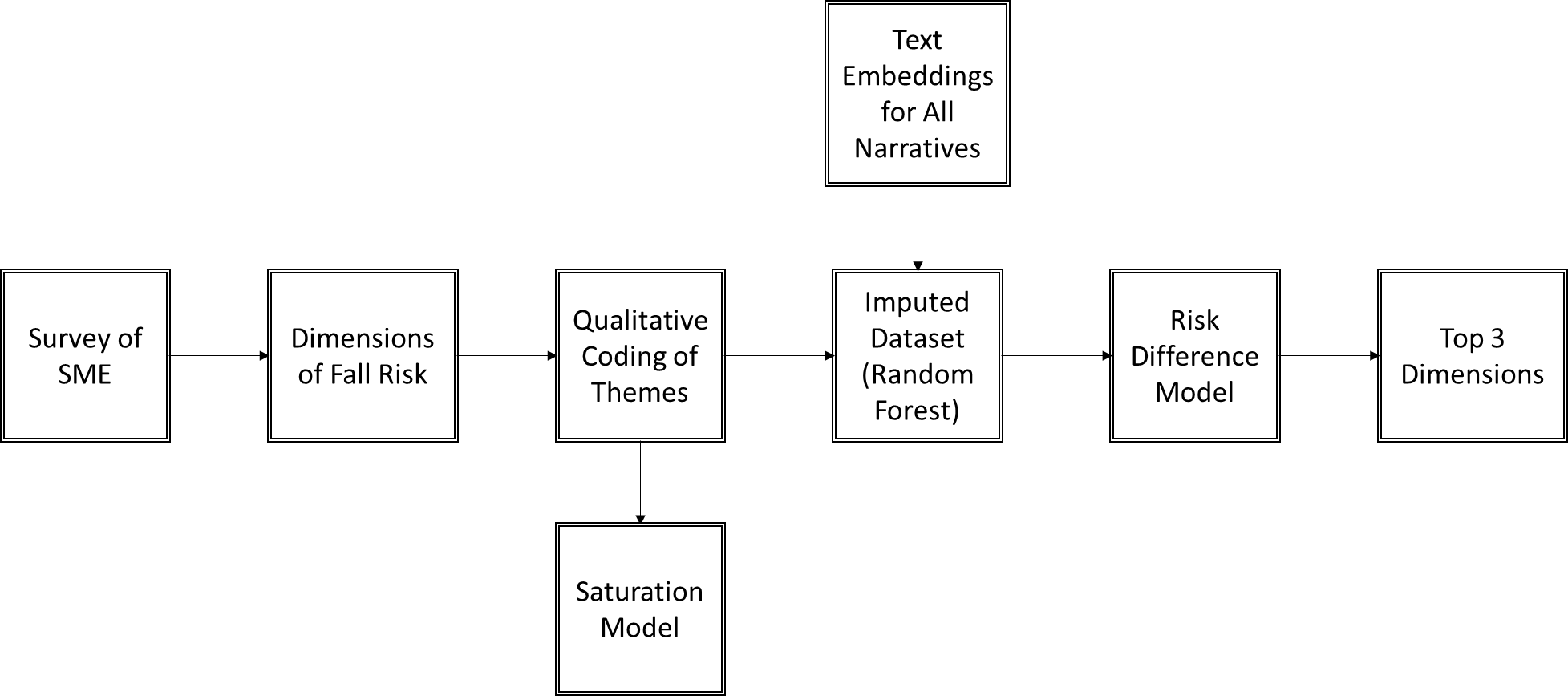
Key findings

* There was a high degree of concordance between the importance of factors identified by clinicians with the results of a natural language processing analysis of NEISS fall narratives.
* Intoxication, loss of consciousness, and patient fragility are the largest contributors to risk of severe fall in our patient cohort. Intoxication and use of intoxicants, in particular, is not identified as a key modifiable fall risk factor in CDC fall prevention guidelines
* Unfamiliar activities and unfamiliar environments were not associated with risk of severe fall, suggesting that elderly individuals should not be discouraged from new activities.

Summary of your approach

Our project uses a unique combination of methodology to develop our model: qualitative analysis, saturation modelling, and NLP. By basing our analysis on qualitative methods, we are able to leverage subject matter expertise and human judgement. This grounds our results in clinical practice and ensures that we create a tool that is acceptable and interpretable to medical providers. However, qualitative methods are labor intensive and not suitable for research on datasets with thousands of records. By applying NLP and random forest imputation, we are able to extend qualitative research to big data applications and develop estimates that are precise and applicable to uncommon subpopulations.



Formally, our project goals were:

* Use machine learning to extend human coding of qualitative research themes to novel data
* Develop a risk reduction tool for falls for use in primary prevention

*Expert Panel*

A panel of practicing physicians was convened to identify factors associated with severe fall risk in the elderly. The panel completed an initial interview, followed by a structured survey. The interview consisted of questions designed to elicit specific themes associated with more severe falls in the elderly. During the survey, panelists ranked 10 identified themes identified both in the interview and in associated literature.

*Qualitative Coding*

Guided by the dimensions identified in the clinical surveys, we performed a deductive thematic analysis of the first 500 narratives. Two individuals manually labeled each narrative with the dimensions described by the narrative and extracted qualitative themes present in each narrative. To ensure consistency, the 500 narratives were double coded, and any discrepancies were resolved by consensus.

*Evaluation of Thematic Saturation*

To ensure that a sufficient number of narratives were coded, we evaluated the thematic saturation of our training sample using methods described in Guest et al, 2020. Briefly, thematic saturation reflects the point where additional data points no longer contribute new useful information to the study. Thematic saturation models apply probability theory to estimate the proportion of total themes that have been described. For our study, we set a baseline of 40% saturation; if any of our themes were below this threshold, we would code an additional 100 narratives.

*Random Forest Imputation*

Once adequate thematic saturation was achieved, we used a random forest model to link the coded themes to text embeddings and apply these relationships to novel data. We created a dataset of indicator variables denoting whether each theme was present in each of the 500 coded narratives. We linked this dataset to the OpenAi text embeddings provided by the contest organizers. Using R's missForest package, we estimated the relationship between the embeddings and our indicator variables and then applied this to the uncoded narratives. We used a model with 100 trees in each forest and 10 variables in each split. We used the first 100 embeddings from the OpenAi model.

*Risk Ratio Model*

We used a negative binomial model on our imputed dataset to estimate the relationship between our themes and risk of severe outcome (Admitted, Held for Observation, or Died). We included all themes in a single model to identify the independent contribution of each. We included age and sex in the model as a potential confounders. We estimated a risk ratio for each theme and selected the top 3 themes for our primary prevention risk reduction tool.

*Negative Binomial Modelling*

We applied the qualitative coding to the remaining 114,628 observations using text embeddings and random forest imputation. A negative binomial model containing all domains and age was derived.

Visualizations

