HHS Public Access

Author manuscript

Int J Med Inform. Author manuscript; available in PMC 2020 October 01.

Published in final edited form as:

Int J Med Inform. 2019 October ; 130: 103943. doi:10.1016/j.ijmedinf.2019.08.003.

Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing

Xin Zhou[a], **Yanshan Wang**[a,*], **Sunghwan Sohn**[a], **Terry M. Therneau**[a], **Hongfang Liu**[a], **David S. Knopman**[b]

[a]Department of Health Science Research, Mayo Clinic, MN, USA

[b]Department of Neurology, Mayo Clinic, MN, USA

Abstract

Introduction: Previous biomedical studies identified many lifestyle exposures that could possibly represent risk factors for dementia in general or dementia due to Alzheimer's disease (AD). These lifestyle exposures are mainly mentioned in free-text electronic health records (EHRs). However, automatic extraction and assessment of these exposures using EHRs remains understudied.

Methods: A natural language processing (NLP) approach was adopted to extract lifestyle exposures and intervention strategies from the clinical notes of 260 patients with clinical diagnoses of AD dementia and 260 age-matched cognitively unimpaired persons. Statistics of lifestyle exposures were compared between these two groups. The mapping results of the NLP extraction were evaluated by comparing the results with data captured independently by clinicians.

Results: Thirty out of fifty-five potentially relevant lifestyle exposures were mentioned in our clinical note dataset. Twenty-two dietary factors and three substance abuses that were potentially relevant were not found in clinical notes. Patients with AD dementia were significantly exposed to more of the potential risk factors compared to the cognitively unimpaired subjects ($\chi^2 = 120.31$, p-value < 0.001). The average accuracy of the automated extraction was 74.0% in comparison with the manual review of randomly selected 50 sample documents.

Discussion and conclusion: We illustrated the feasibility of NLP techniques for the automated evaluation of a large number lifestyle habits using free-text EHR data. We found that AD dementia patients were exposed to more of the potential risk factors than the comparison

*Corresponding author. wang.yanshan@mayo.edu (Y. Wang).

Author statement

Study conception and design: XZ, YW, HL, DK.
Acquisition of data: XZ, YW, TT.
Analysis and interpretation of data: XZ, YW, SS, DK.
Drafting of manuscript: XZ, YW.
Critical revision: XZ, YW, SS, HL, DK.
Final Approval: XZ, YW, SS, TT, HL, DK.
All authors have seen and approved the final version of the manuscript being submitted. They warrant that the article is the authors' original work, hasn't received prior publication and isn't under consideration for publication elsewhere.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2019.08.003.

group. Our results also demonstrated the feasibility and accuracy of investigating putative risk factors using NLP techniques.

Keywords

Alzheimer's disease; Electronic health records; Natural language processing; Lifestyle exposure

1. Introduction

Alzheimer's disease (AD) is a degenerative brain disease and the most common form of dementia in the United States [1]. Alzheimer's disease affects about 5.7 million Americans and is growing exponentially with an estimate of 13.8 million being affected by midcentury [2]. Since the number of patients with AD is expected to grow, finding ways to prevent and lower the risk of AD is crucial [3]. The majority of investigations into potential lifestyle-related risk factors for AD dementia are retrospective cross-sectional analyses based on selfreport questionnaires [4]. However, questionnaire assessment is limited to patient's subjective experience, which is not suitable for objective measurements such as laboratory test results [5]. Moreover, many questionnaires investigate the exposure factors, such as food intake over one year instead of daily records or 24 -h recalls, resulting in an increased recall bias [6]. Furthermore, the exposures studied in questionnaire surveys are not scalable. Table 1 lists the recent lifestyle risk factor studies [7-11] of AD, in which only 1–3 exposures were investigated in each survey, highlighting the need for a more comprehensive approach which would produce more valid and meaningful results.

The growing availability of electronic health records (EHRs) provides an increased opportunity for a thorough investigation of risk factors for AD dementia. EHRs refer to the comprehensive records of a patient health care history that resides in digital format [12,13]. Clinical notes are free-text EHRs that contain textual descriptions of physician-patient encounters and capture the information that the author intended to collect concerning a certain medical topic, offering valuable resources for identifying lifestyle exposures that physicians believed to be clinically important [14,15]. However, since clinical notes are free-text narratives lacking a standardized structure, searching for simple keywords may result in low sensitivity [16,17]. Natural language processing (NLP) offers a solution for clinical notes processing. NLP is a field of computational techniques that allows computers to extract relevant information from human language [18] and offers a viable solution for effectively processing clinical notes. It has been widely utilized in clinical applications, such as quality measurement of laboratory tests [19], early diagnosis of disease [20], and suicide behavior screening [21]; however, to best of our knowledge, none of the previous NLP approaches have been applied to lifestyle investigation.

This study sought to apply NLP techniques to automatically extract and assess lifestyle exposures as well as corresponding intervention strategies from the EHRs of AD patients and non-demented controls. Our study was motivated by the needs to: 1) demonstrate the feasibility of NLP techniques in lifestyle investigation for AD patients; 2) determine whether previously identified lifestyle exposures were recognized in primary care settings; and 3) determine whether lifestyle interventions were delivered by primary care providers. Our

dataset of clinical notes has been an invaluable resource to allow examination of lifestyle assessments and management among AD patients in primary health care settings.

2. Methods

2.1. Study subjects

The patient cohort was filtered from the Mayo Clinic Employee and Community Health (ECH), a department which delivers primary care to 140,000 patients who reside within the area surrounding Rochester, Minnesota [22]. The inclusion criteria of the cohort were: 1) a patient who received the primary care in Mayo Clinic, Rochester from 1998 to 2015; 2) a patient whose EHR data was available; 3) the presence of research authorization for using medical records for research. Patients were considered to have AD dementia based on being assigned ICD-9-CM code 331.0 (Alzheimer's disease). A control group of patients were randomly selected from the ECH cohort who matched the average age of the AD group and did not have any ICD diagnosis code for AD or dementia. A total of 260 AD dementia patients (mean age = 82.28 ± 9.47) and 260 age-matched controls (mean age = 81.05 ± 7.80) were identified. All of their clinical notes (64,672 in total) were extracted from the clinical data warehouse and the documents of each subject were merged into a single clinical note for further analysis. This study was approved by the Institutional Review Board (IRB) for human subject research.

2.2. Study design

This case-control study aims to investigate the prevalence of lifestyle risk factor exposures among AD dementia patients in primary care settings. A schematic diagram of the study is presented in Fig. 1. First, we retrieved clinical notes of AD dementia patients and age-matched controls. Processing clinical notes requires tools that identify standard medical concepts from free-form text [23]. MetaMap (https://metamap.nlm.nih.gov/) is such a tool that automatically maps biomedical texts to standard medical concepts in the Unified Medical Language System (UMLS) [24-27]. MetaMap shows comparable or better performances to other mapping tools such as clinical Text Analysis and Knowledge Extraction System (cTAKES) [28,29]. We chose MetaMap since it is simpler to implement than cTAKES for people without a Java programming background.

We utilized 55 lifestyle risk factors that have been identified by Kostoffa et al (Table S1) [30] as potentially being related to the development of AD dementia. These factors were retrieved from the AD-related literature consisting of 100,000 Medline abstracts and 99,610 PubMed articles. Then we collected all UMLS concepts related to each risk factor (e.g., *vitamin B deficiency*) and its corresponding intervention strategy (e.g., *vitamin B supplement)* by using online UMLS Terminology Services (default settings) as the searching browser (https://www.nlm.nih.gov/research/umls/). A UMLS concept dictionary of lifestyle exposures and a dictionary of interventions were established (described below). By comparing UMLS concepts extracted from clinical notes with two dictionaries, we can identify whether the lifestyle exposures as well as intervention strategies occurred in clinical notes.

2.3. Dictionary of lifestyle exposures

To construct the dictionary, we searched for all of the UMLS concepts corresponding to lifestyle exposure terms from UMLS browser. Searching keywords of each exposure factor were determined as following: 1) for excessive/deficient dietary factors, searching keywords are the same as the factors (e.g., *"high fat diet"* are the keywords for *high fat diet);* 2) for food additives, searching keywords are *"exposure factor + diet"* (e.g., *"Menadione diet"* for the factor *Menadione*); 3) for substance abuse, keywords are *"exposure factors + abuse"* (e.g., *"amphetamine abuse"* for the factor *amphetamine).* All searching keywords were shown in the appendix Table S1.

2.4. Dictionary of intervention strategies

To construct the dictionary, we searched for all of the UMLS concepts corresponding to lifestyle intervention terms from UMLS browser. Searching keywords were determined as following: 1) for excessive dietary factors such as *high ** diet,* searching keywords are *"low ** diet"* (e.g., *"low fat diet"* are the keywords for *high fat diet);* 2) for dietary deficiencies such as ** *deficiency,* we search "** *supplement"* (e.g., *"vitamin B supplement"* are the keywords for *vitamin B deficiency);* 3) *"low + exposure factor + diet"* are the keywords for food additives (e.g., *"low cysteine diet"* are the keywords for *cysteine);* 4) intervention keywords for *sleep disorder* include all benzodiazepine drugs that are marketed in USA, which are presently the most frequently prescribed hypnotics [31].

2.5. Extracting lifestyle exposures and interventions using NLP

NLP is intended to take the place of manual chart review and extract information from texts automatically. There are three steps in the process, as shown in Fig. 2.

Step 1: Mapping clinical notes to UMLS concepts.—We applied the default setting of MetaMap Windows version to index UMLS concepts from clinical notes (n = 260 in each group). Fuzzy mapping was not applied in our approach since fuzzy mapping has the potential to cause many unrelated mapping results.

Step 2: Extracting lifestyle exposures from clinical notes.—We investigated the occurrence of lifestyle exposure-related UMLS concepts among mapped clinical notes. For factors mentioned in clinical notes, number of records (denoted as m), number of patients (denoted as n) and percentage of patients were recorded. Number of records was defined as the total frequency of the certain exposure in 260 clinical notes. Number of patients was defined as the number of clinical notes that contained the specific exposure. Percentage of patients = number of patients/260. In addition, the total number of exposures was defined as how many different lifestyle exposures each patient had.

Step 3: Extracting lifestyle interventions strategies from AD sections and the complete clinical notes.—When extracting intervention strategies from clinical notes, we aimed to figure out whether these strategies were related to AD or other diseases. We used the Python NLTK package to filter clinical note sections that only record AD-related information only. As shown in Fig. 3, for patients with multiple diseases, each diagnosis and the related medical information (such as disease descriptions, medications, doctors' advices,

etc.) were recorded separately. Medical records following the diagnosis of AD were considered as AD sections. Intervention strategies in AD sections were extracted, and reflected physicians' attitudes and practices toward prevention and treatment of AD. Next, we expanded our searching range to the whole corpus of clinical notes and extracted all lifestyle interventions. Number of exposures, number of patients, and percentage of patients were recorded.

2.6. Limited evaluation of the NLP method in MetaMap

We randomly sampled 50 documents from the dataset for manual review. Both lifestyle risk factors and intervention strategies were manually annotated by investigators and this method was considered as the “gold standard”. Because this study does not focus on the evaluation of NLP methods in MetaMap, when MetaMap found a matching UMLS map, we manually evaluated how accurate the mapping was. Findings from the automated algorithm were compared with the gold standard, and the precision of the NLP method was calculated.

2.7. Statistical analysis

All 520 subjects were included in the statistical analysis. We compared the demographic characteristics between AD dementia patients and control subjects using Pearson’s χ^2 test for categorical variables and Student’s *t*-test for continuous variables. Generalized linear models were fit to obtain the odds ratio (OR) and 95% confidence intervals of the top 10 exposure factors that occurred in clinical notes. All statistical analyses were performed using R statistical software (version 3.5.1, R Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1. Study cohort demographics

The study sample consisted of 260 AD dementia patients and 260 controls from a general primary care population. Their demographics are summarized in Table 2. There was no significant difference in age, sex, race, and marital status found between AD dementia patients and the controls. Compared with the general population in the United States, this study cohort had a higher proportion of white race in both groups (97.3% in AD group, 99.2% in control group vs. 62.0% in the general population) with a correspondingly lower proportion in other races.

3.2. Lifestyle exposures in AD patients

Using NLP techniques, we identified 20 out of 55 lifestyle risk factor exposures from clinical notes among patients with AD dementia, which could be categorized as dietary factors, daily activity and substance abuse (Table 3). Overall, *tobacco smoking* and *malnutrition* were the most common exposure factors among AD patients, affecting 145 and 134 patients respectively. 7 categories of vitamin or mineral deficiencies (including *vitamin B/D/E deficiency* and *potassium/iron/magnesium/calcium deficiency*) were also identified in our cohort. Conversely, cardiovascular/metabolic exposures, such as *high fat diet, high calorie diet, high carbohydrate diet,* etc., were relatively rare in this cohort ($n < 3$).

3.3. Lifestyle interventions in AD dementia patients

Lifestyle intervention strategies retrieved from AD sections and whole clinical notes are compared in Table 4. The intervention records were very rare in AD sections with only five intervention strategies being suggested by physicians, such as *physical activity* (n = 123), *cognitive activity* (n = 45), *dietary supplement* (n = 19), *low salt diet* (n = 1) and *benzodiazepines* (n = 1). However, when expanding our searching range to the whole clinical notes, a total of 23 lifestyle intervention strategies were identified. *Physical activity* was the most commonly identified lifestyle intervention suggested by physicians (n = 227), followed by *dietary supplement* (n = 148), *smoking sessation* (n = 137), *fish oil supplement* (n = 94) and vegetables *supplement* (n = 89). In addition, vitamins (including vitamin B, C, D, and E) and mineral supplements (including potassium, iron, zinc, magnesium and calcium) were widely encouraged by physicians.

3.4. Lifestyle exposures between AD and the control groups

We identified 13 lifestyle exposures from the clinical notes of the control group, (Table 5) with most exposures (11 of 13, except *vitamin K deficiency* and *high iron diet)* being found in the AD group. Individual effects and overall effects of lifestyle exposures were investigated between AD patients and non-demented controls.

3.4.1. Individual effect—We analyzed the individual effects of the 10 lifestyle exposures that occurred the most frequently in the clinical notes, which were the same between the AD dementia and control groups. Fig. 4 provides the results from both univariable analysis and multivariable analysis. In univariable analysis, after adjustment for age and sex, all 10 exposures were significantly associated with AD dementia, except *iron deficiency* (adjusted OR: 1.59, 95% CI: (0.98, 2.58)). In the multivariable analysis, after adjustments for age, sex, and each of the other 9 exposures, 5 out of the 10 factors were significantly associated with AD dementia. *Potassium deficiency* showed the strongest correlation with AD (adjusted OR: 3.94, 95% CI: (1.63, 10.64)) followed by *calcium deficiency* (adjusted OR = 3.54, 95% CI: (1.43, 9.87)) and *tobacco smoking* (adjusted OR = 3.44, 95% CI: (2.23, 5.33)). *Malnutrition* (adjusted OR = 2.78, 95% CI: (1.74, 4.48)) and *excess alcohol use* (adjusted OR = 2.23, 95% CI: (1.31, 3.82)) were also associated with AD dementia, but the associations were weaker than the aforementioned factors. We did not identify any associations between the AD dementia and the other 5 exposures, including *iron deficiency* (adjusted OR: 0.65, 95% CI: (0.35, 1.19)), *dehydration* (adjusted OR: 1.28, 95% CI: (0.74, 2.19)), *vitamin D deficiency* (adjusted OR: 1.78, 95% CI: (0.87, 3.74)), *vitamin B deficiency* (adjusted OR: 2.70, 95% CI: (0.99, 8.67)) and *magnesium deficiency* (adjusted OR: 3.14, 95% CI: (0.83, 13.53)). The effects of lifestyle exposures outside of the top 10 factors (e.g., *high fat diet*) were not analyzed due to their low frequencies (number of patient ≤ 1 in each group).

Summarizing all lifestyle exposures, we observed significantly more risk factors in the AD dementia group compared to the control group ($\chi^2 = 120.31$, p-value < 0.001). The median of the total number of exposures each patient had was 3 (interquartile range: 1–5) in the AD group whereas the number was 1 (interquartile range: 0–2) in the control group. The distribution of lifestyle exposures between two groups is compared in Fig. 5. Note that we

excluded an outlier patient in the AD group who had as much as 20 exposures in total. Fig. 6 shows the odds of AD according to the number of lifestyle exposures. A strong and graded relation was noted between the number of exposures and the presence of AD, with an odds ratio of 13.2 for top (number of exposures = 6) versus the lowest decile of exposure numbers (number of exposures = 0).

3.5. Missing lifestyle exposures

25 out of 55 potential lifestyle exposures identified by literature review were not identified in our dataset, 22 of which were dietary factors (Table 6). 13 dietary factors could not be mapped to any UMLS concept, including dietary excesses (*diabetogenic diet, high advanced glycation end products diet, high arachidonic acid,* and *high unfermented soy),* nutrient deficiencies (*linoleic acid deficiency, early life nutrient restriction, glutathione depletion, low cocoa, low coffee, low flavonols* and *low fruit),* and food additives (*menadione. diacetyl).* The other 9 factors could be mapped to UMLS concepts but were not mentioned in clinical notes, including dietary heavy metal intake *(high copper diet, high zinc diet),* food additives *(monosodium glutamate, cysteine),* nutrient deficiencies *(selenium deficiency, glucose deprivation, low tryptophan diet), high methionine diet* and *industrialized/preserved food.* 3 substance abuses were not found in our patient cohorts, including *amphetamine, MDMA,* and *cocaine/opiates.* A full list of the 55 lifestyle exposures not found in our dataset is provided in Table S1 of the supplementary material.

3.6. Performance of the NLP method in MetaMap

The performance of the automated NLP algorithm was compared with the manual review of the random 50 sample documents. Specifically, 73 medical concepts related to risk exposures and interventions were found by MetaMap in the 50 sampled documents, among which the true positive is 54 and the false positive is 19. Thus, the precision of MetaMap is 74.0% based on our limited manual evaluation. Detailed information about the results in terms of frequency, number of true positives, and PPV of each extracted concept and a few typical false positive examples are provided in Tables S2 and S3 of the supplementary material, respectively.

4. Discussion

This is, to the best of our knowledge, the first study that examined the published lifestyle exposures and corresponding intervention strategies for AD patients in routine practice using NLP techniques. By investigating 55 lifestyle exposure factors from 520 clinical notes comprised of 64,672 medical documents, we found evidence of a positive graded relation between AD dementia and the number of lifestyle exposures the patients suffered. The results in this exploratory study show that NLP techniques are able to access much more lifestyle information than could be obtained in commonly used questionnaire assessments for this task. The ability to evaluate large numbers of potential risk factors illustrates the feasibility of NLP techniques for lifestyle risk factor evaluation in a large-scale cross-sectional study.

The 55 lifestyle risk factors investigated in this study were retrieved from different levels of AD-related studies such as neuronal cell culture [32], transgenic animal models [33], brain imaging studies [34], etc. However, only a few of them had been proven to significantly increase the risk of AD in well-designed longitudinal clinical studies. In our dataset, 5 lifestyle exposures were significantly related to the presence of AD, including *potassium deficiency, calcium deficiency, tobacco smoking, malnutrition* and *excessive alcohol use.* However, a causal relation between these factors and the onset of AD could not be determined in this cross-sectional study. Large-scale clinical trials should be conducted in the future to further investigate their roles in AD progression. 22 dietary factors were not mentioned in our clinical notes datasets. It is beyond the scope of this work to consider whether all putative risk factors ought to be reported in clinic notes, or used for disease monitoring.

Most current human studies of lifestyle exposures of AD focus on cardiovascular exposures, such as *high fat diet* [35], *high calorie diet* [36], and *high carbohydrate diet* [37]. These factors attract much attention due to their high correlations with metabolic and cardiovascular diseases including diabetes, hypertension, stroke, etc. However, we found that cardiovascular exposures were rarely observed in the clinical notes of the AD cohort. For example, *high calorie diet* and *high meat diet* occurred in the clinical notes of only 1 and 2 patients out of 260 patients, respectively. In contrast, nutrient deficiency exposures occurred much more commonly occurred in AD patients' clinical notes. Over 25 patients were exposed to each of the following: 1) *vitamin B deficiency;* 2) *vitamin D deficiency;* 3) *potassium deficiency;* 4) *iron deficiency;* 5) *calcium deficiency* and 6) *malnutrition. Vitamin E deficiency, magnesium deficiency* and *starvation* were also found in the clinical notes of our patient cohort. Our findings are in line with previous studies reporting vitamin and mineral insufficiency in patients with AD [38-41]. These studies have investigated the risk or the prevalence of certain nutrient deficiency in AD patients in specific contexts such as in elder AD patients, in elder women, or in the general population. Our study adds to previous findings by reporting the prevalence of nutrient deficiencies 1) using NLP methods to perform complete nutrient assessments, and 2) in other patient groups.

Nutrient supplements were widely identified in our clinical note datasets where we found 14 nutrient supplements, covering vitamins, minerals, vegetables, docosahexaenoic acids and fatty acids. However, the laboratory test results of the circulating concentrations of nutrients were rarely identified in our dataset. For example, among 21 AD patients who took vitamin D supplements in our cohort, only 1 of them had the vitamin D measurements in his medical records. Whether the other 20 patients had satisfactory circulating levels of vitamin D after receiving vitamin D therapy was not mentioned in their clinical notes. In addition, previous studies suggested a remarkably low adherence of nutrition supplements. Miller et al [42] reported a poor adherence (67%) with nutrition supplement prescriptions over 42 days. Modi et al [43] reported a lower adherence (30%) to prescribed multivitamin therapy in patients over six months. Hayes et al [44] reported that the medication adherence was significantly lower among subjects over 65 with cognitive impairments. Finally, some nutrients (e.g., *selenium deficiency)* were never routinely measured in our dataset. Thus, the intervention strategies mentioned in clinical notes may not cover all nutrient components.

Most previous lifestyle studies of AD focused on a limited number of lifestyle exposures in each survey, thus the synergistic effects or the overall effects of multiple lifestyle habits could not be analyzed. By comparing the distributions of lifestyle exposures between the clinical notes from the AD cohort and those from the control cohort, we found that the clinical notes from the AD group had recorded significantly more exposures than those from the control. This finding confirms the results about previously-studied lifestyle exposures. A positive graded relation was observed between AD dementia and the number of exposures the patient had, without either a threshold or a plateau. In particular, even a single lifestyle exposure is significantly related to AD, suggested the importance of complete lifestyle assessment and modification. AD and many other health consequences might share common lifestyle risk factors with overlapping pathological processes. Martins et al [45] found that abnormal glycolipid metabolism are indicated as central in the pathogenesis of both diabetes and AD. Wells et al [46] demonstrated that nutritional deficiencies in elder patients were strongly associated with cognitive impairment and vascular diseases. Hence monitoring a single risk factor may lead to many health benefits.

MetaMap for extracting information from clinical notes has been evaluated in the previous studies and shown high precision and recall [47]. However, the irregular expressions, abbreviations or misspellings contained in free-text clinical notes might result in misrecognitions. For example, the intervention strategy of *vitamin D supplement* could not be retrieved from the sentence “She is on calcium carbonate and D”. In addition, MetaMap could not distinguish lifestyle risk factors and interventions. *High fat diet* could be considered as 1) an unhealthy diet habit, and 2) a special diet regimen suitable for patients who need such nutrition, though the second situation was very rare. Finally, although most negative context could be successfully mapped to negative UMLS terms, some of them might cause false positive results. For example, both “fat free” and “non-fat diet” could be successfully mapped to negative UMLS terms while “not high in fat” was falsely mapped to *UMLS Term C0425441- Diet high in saturated fats.* Based on the limited evaluation on the random samples, the average accuracy of MetaMap in identifying UMLS concepts associated with lifestyle exposures and interventions is 74.0%.

There are several limitations in our study that need to be considered when interpreting the results. First, we did not develop comprehensive algorithms to map UMLS concepts and account for lifestyle exposure terms that could not be mapped to UMLS concepts. Several dietary exposures with complex expressions could not be directly mapped to any UMLS concepts, such as *early life nutrient restriction* and *high advanced glycation end products diet.* A keywords-based search could be an alternative approach for investigation of these exposures in a future study. This research was a pilot study to investigate the potential of using NLP algorithm to extract lifestyle exposures from clinical notes for clinical research. In future work, we would like to develop a more comprehensive algorithm that could identify the medical concepts that are related but linguistically different, and improve the NLP method using more heuristic rules or leveraging machine learning approaches to find lifestyle exposures that could not be mapped to UMLS concepts. Second, we did not have readily available data on patients’ education level from EHRs. Therefore, the association between the presence of AD and exposure factors may reflect disturbances associated with socioeconomic or education level. Third, the length of medical records varies among

patients. For patients who lived in Rochester, MN for a short time and only had records for a few visits, lifestyle information may not be completely recorded in their clinical documents. The results may also be biased due to the difference between patients and controls in the number of visits. The average number of visits for each patient in the case group is 214 while this number is 35 for the control group. During the control selection, we did not match the number of visits or the number of notes between the case and control. The reason is that the lifestyle information is usually recorded as part of clinical notes during any patient's visit. Having said that, the patients with more visits may possibly generate more information about lifestyle exposures than those in the control, which may introduce bias to this study. Finally, we cannot confirm that these lifestyle exposures are risk factors leading to AD since we included all clinical notes of patients, involving visits before and after AD diagnosis for studying both risk factors and interventions. It is also possible that AD leads to these lifestyle exposures, such as potassium deficiency and calcium deficiency. In future work, we will extract risk factors and intervention information from notes of visits before and after AD diagnosis, respectively.

5. Conclusion

The results in this exploratory study illustrate the feasibility of NLP techniques for evaluating multiple lifestyle risk factors in a large-scale cross-sectional study by using EHRs. In contrast to contemporary questionnaire-based lifestyle investigations which missed the risk factors that are outside the scope of research interests, our novel NLP approach allows a complete lifestyle assessment in an efficient and costeffective manner. With an accurate automatic method for investigating lifestyle exposures, larger-scale epidemiologic research could be feasibly conducted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH grants UL1TR02377 Supplement, U01TR002062, and R01LM11934. The funders had no role in the design of the study, or collection, analysis, and interpretation of data and in preparation of the manuscript. The views presented in this report are not necessarily representative of the funder's views and belong solely to the authors.

References

[1]. Yankner BA, Mechanisms of neuronal degeneration in Alzheimer's disease, Neuron 16 (5) (1996) 921–932.

[2]. Association, A.s, Alzheimer's Disease Facts and Figures, Available from: (2018) https://www.alz.org/alzheimers-dementia/facts-figures.

[3]. Arab L, Sabbagh MN, Are certain lifestyle habits associated with lower Alzheimer's disease risk? J. Alzheimers Dis. 20 (3) (2010) 785–794. [PubMed: 20182018]

[4]. van Duijn CM, Epidemiology of the dementias: recent developments and new approaches, J. Neurol. Neurosurg. Psychiatry 60 (5) (1996) 478–488. [PubMed: 8778250]

[5]. Launer LJ, et al., Rates and risk factors for dementia and Alzheimer's disease: results from EURODEM pooled analyses. EURODEM Incidence Research Group and Work Groups. European Studies of Dementia, Neurology 52 (1) (1999) 78–84. [PubMed: 9921852]

[6]. Sempos CT, Invited commentary - some limitations of semiquantitative food frequency questionnaires, Am. J. Epidemiol. 135 (10) (1992) 1127–1132.

[7]. Berti V, et al., Nutrient patterns and brain biomarkers of Alzheimer's disease in cognitively normal individuals, J. Nutr. Health Aging 19 (4) (2015) 413–423. [PubMed: 25809805]

[8]. Gardener S, et al., Adherence to a Mediterranean diet and Alzheimer's disease risk in an Australian population, Transl. Psychiatry 2 (2012) el64.

[9]. Gauthier E, et al., Environmental pesticide exposure as a risk factor for Alzheimer's disease: a case-control study, Environ. Res. 86 (1) (2001) 37–45. [PubMed: 11386739]

[10]. Mosconi L, et al., Lifestyle and vascular risk effects on MRI-based biomarkers of Alzheimer's disease: a cross-sectional study of middle-aged adults from the broader New York City area, BMJ Open 8 (3) (2018) p. e019362.

[11]. Rogers MA, Simon DG, A preliminary study of dietary aluminium intake and risk of Alzheimer's disease, Age Ageing 28 (2) (1999) 205–209. [PubMed: 10350420]

[12]. Kim EH, et al., Challenges to using an electronic personal health record by a lowincome elderly population, J. Med. Internet Res. 11 (4) (2009) e44. [PubMed: 19861298]

[13]. Wang Y, et al., Clinical information extraction applications: a literature review, J. Biomed. Inform. 77 (2018) 34–49. [PubMed: 29162496]

[14]. Kipper-Schuler Karin, James Masanz VK, Ogren Philip, Savova Guergana, System evaluation on a named entity corpus from clinical notes, Language Resources and Evaluation Conference, LREC, (2008).

[15]. Meng F, et al., Automatic generation of repeated patient information for tailoring clinical notes, Int. J. Med. Inform. 74 (7–8) (2005) 663–673. [PubMed: 16043089]

[16]. Cao H, Stetson P, Hripcsak G, Assessing explicit error reporting in the narrative electronic medical record using keyword searching, J. Biomed. Inform. 36 (1–2) (2003) 99–105. [PubMed: 14552851]

[17]. Hripcsak G, et al., Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports, Radiology 224 (1) (2002) 157–163. [PubMed: 12091676]

[18]. Cambria E, White B, Jumping NLP curves: a review of natural language processing research, IEEE Comput. Intell. Mag. 9 (2) (2014) 48–57.

[19]. Mehrotra A, et al., Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures, Gastrointest. Endosc. 75 (6) (2012) p. 1233–9 el4. [PubMed: 22482913]

[20]. Kaur H, et al., Automated chart review utilizing natural language processing algorithm for asthma predictive index, BMC Pulm. Med. 18 (1) (2018) 34. [PubMed: 29439692]

[21]. Zhong QY, et al., Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing, BMC Med. Inform. Decis. Mak. 18 (1) (2018) 30. [PubMed: 29843698]

[22]. DeJesus RS, Use of a clinical decision support system to increase osteoporosis screening, J. Eval. Clin. Pract. 18 (4) (2012) 926. [PubMed: 22747584]

[23]. Pratt W, Yetisgen-Yildiz M, A study of biomedical concept identification: MetaMap vs. people, AMI A Annu Symp Proc, (2003), pp. 529–533.

[24]. Aronson AR, Lang FM, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236. [PubMed: 20442139]

[25]. Bodenreider O, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) D267–70 (Database issue). [PubMed: 14681409]

[26]. Meystre S, Haug PJ, Evaluation of medical problem extraction from electronic clinical documents using MetaMap transfer (MMTx), Stud. Health Technol. Inform. 116 (2005) 823–828. [PubMed: 16160360]

[27]. Shah NH, et al., Comparison of concept recognizers for building the Open Biomedical Annotator, BMC Bioinformatics 10 Suppl 9 (2009) S14.

[28]. Reátegui R, Ratté S, Comparison of MetaMap and cTAKES for entity extraction in clinical notes, BMC Med. Inform. Decis. Mak. 18 (3) (2018) 74. [PubMed: 30255810]

[29]. Wu Y, et al., A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries, AMIA Annual Symposium Proceedings (2012).

[30]. Kostoff RN, Porter AL, Buchtel HA, Prevention and Reversal of Alzheimer's Disease: Treatment Protocol, (2018).

[31]. Monti JM, Sleep laboratory and clinical-studies of the effects of triazolam, flunitrazepam and flurazepam in insomniac patients, Methods Find. Exp. Clin. Pharmacol. 3 (5) (1981) 303–326. [PubMed: 6120270]

[32]. Kawahara M, et al., Alzheimer's beta-amyloid, human islet amylin, and prion protein fragment evoke intracellular free calcium elevations by a common mechanism in a hypothalamic GnRH neuronal cell line, J. Biol. Chem. 275 (19) (2000) 14077–14083. [PubMed: 10799482]

[33]. Sensi SL, et al., Altered oxidant-mediated intraneuronal zinc mobilization in a triple transgenic mouse model of Alzheimer's disease, Exp. Gerontol. 43 (5) (2008) 488–492. [PubMed: 18068923]

[34]. Saar G, et al., Laminar specific detection of APP induced neurodegeneration and recovery using MEMRI in an olfactory based Alzheimer's disease mouse model, Neuroimage 118 (2015) 183–192. [PubMed: 26021215]

[35]. Laitinen MH, et al., Fat intake at midlife and risk of dementia and Alzheimer's disease: a population-based study, Dement. Geriatr. Cogn. Disord. 22 (1) (2006) 99–107. [PubMed: 16710090]

[36]. Luchsinger JA, et al., Caloric intake and the risk of Alzheimer disease, Arch. Neurol. 59 (8) (2002) 1258–1263. [PubMed: 12164721]

[37]. Young KW, et al., A randomized, crossover trial of high-carbohydrate foods in nursing home residents with Alzheimer's disease: associations among intervention response, body mass index, and behavioral and cognitive function, J. Gerontol. A Biol. Sci. Med. Sci. 60 (8) (2005) 1039–1045. [PubMed: 16127110]

[38]. Brewer GJ, et al., Subclinical zinc deficiency in Alzheimer's disease and Parkinson's disease, Am. J. Alzheimers Dis. Other Demen. 25 (7) (2010) 572–575. [PubMed: 20841345]

[39]. Clarke R, et al., Folate, vitamin B12, and serum total homocysteine levels in confirmed Alzheimer disease, Arch. Neurol. 55 (11) (1998) 1449–1455. [PubMed: 9823829]

[40]. Littlejohns TJ, et al., Vitamin D and the risk of dementia and Alzheimer disease, Neurology 83 (10) (2014) 920–928. [PubMed: 25098535]

[41]. Sato Y, Asoh T, Oizumi K, High prevalence of vitamin D deficiency and reduced bone mass in elderly women with Alzheimer's disease, Bone 23 (6) (1998) 555–557. [PubMed: 9855465]

[42]. Miller MD, et al., Adherence to nutrition supplements among patients with a fallrelated lower limb fracture, Nutr. Clin. Pract. 20 (5) (2005) 569–578. [PubMed: 16207699]

[43]. Modi AC, et al., Adherence to vitamin supplementation following adolescent bariatric surgery, Obesity (Silver Spring) 21 (3) (2013) E190–5. [PubMed: 23404956]

[44]. Hayes TL, et al., Medication adherence in healthy elders: small cognitive changes make a big difference, J. Aging Health 21 (4) (2009) 567–580. [PubMed: 19339680]

[45]. Martins I, et al., Apolipoprotein E, cholesterol metabolism, diabetes, and the convergence of risk factors for Alzheimer's disease and cardiovascular disease, Mol. Psychiatry 11 (8) (2006) 721. [PubMed: 16786033]

[46]. Wells JL, Dumbrell AC, Nutrition and aging: assessment and treatment of compromised nutritional status in frail elderly patients, Clin. Interv. Aging 1 (1) (2006) 67. [PubMed: 18047259]

[47]. Reategui R, Ratte S, Comparison of MetaMap and cTAKES for entity extraction in clinical notes, BMC Med. Inform. Decis. Mak. 18 (2018).

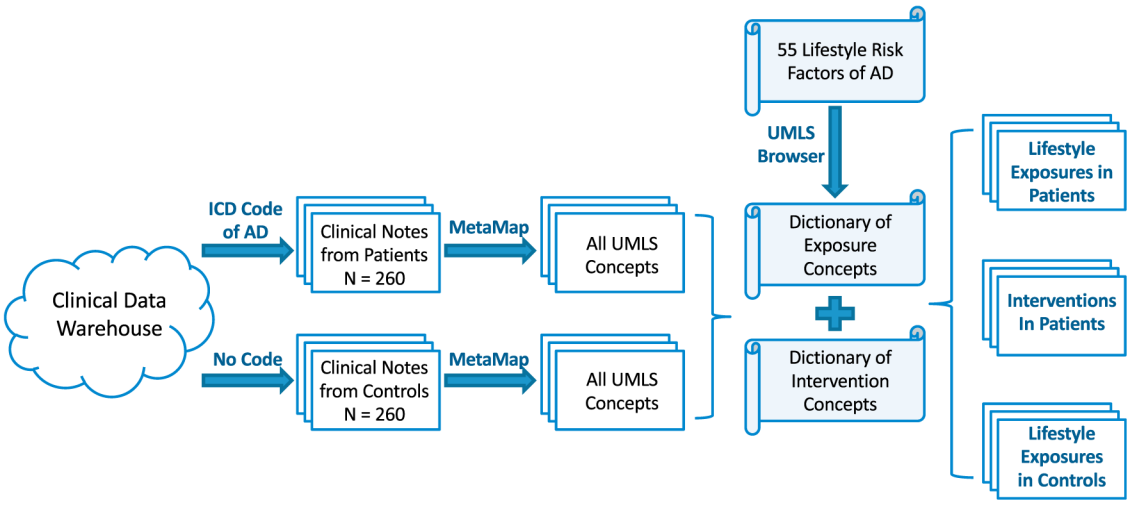


Fig. 1.
Procedure for retrieving lifestyle information from clinical notes.

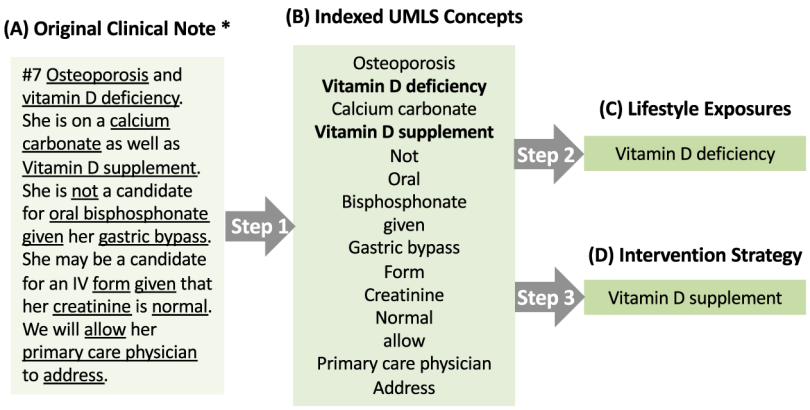


Fig. 2.

Extracting lifestyle exposures and modifications from clinical notes. * This is not an AD section. Step 1: indexing UMLS concepts from original clinical notes. Step 2: identifying lifestyle exposures. Step 3: identifying lifestyle intervention strategies.

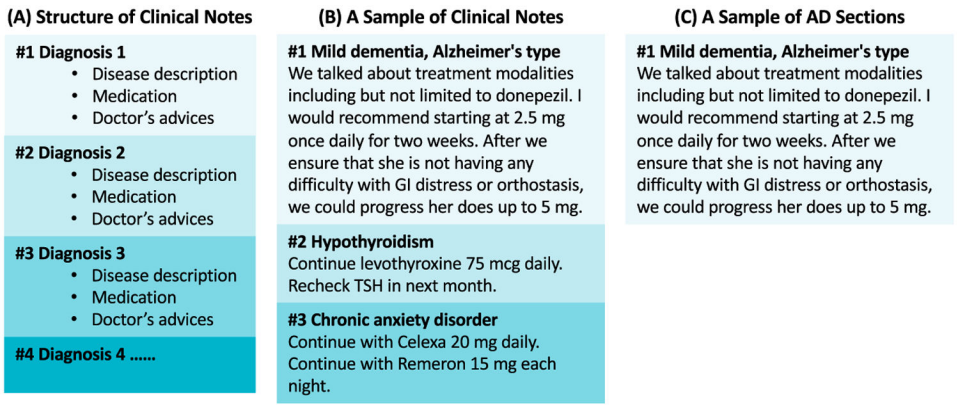
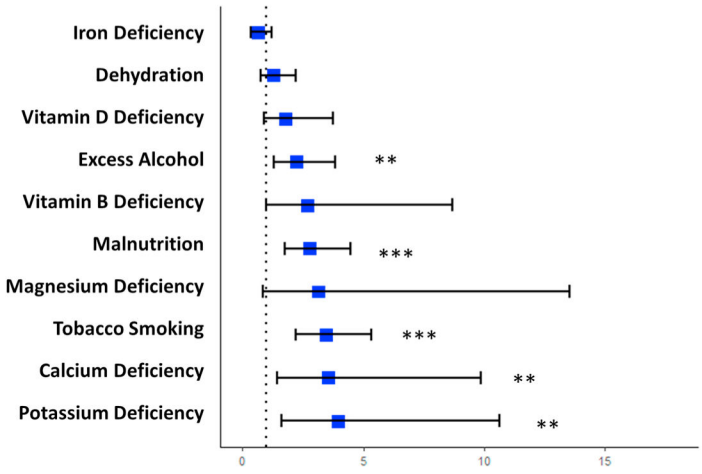


Fig. 3.
Illustration of extracting AD-related sections from a synthetic clinical note.



	Univariable analysis		Multivariable analysis	
	Adjusted OR	P-value	Adjusted OR	P-value
Iron Deficiency	1.59 (0.98, 2.58)	0.060	0.65 (0.35, 1.19)	0.167
Dehydration	3.25 (2.12, 5.06)	< 0.001	1.28 (0.74, 2.19)	0.372
Vitamin D Deficiency	3.16 (1.77, 5.92)	< 0.001	1.78 (0.87, 3.74)	0.119
Excess Alcohol	3.44 (2.20, 5.49)	< 0.001	2.23 (1.31, 3.82)	0.003
Vitamin B Deficiency	6.60 (2.74, 19.66)	< 0.001	2.70 (0.99, 8.67)	0.067
Malnutrition	5.10 (3.45, 7.63)	< 0.001	2.78 (1.74, 4.48)	< 0.001
Magnesium Deficiency	3.42 (1.20, 12.23)	0.033	3.14 (0.83, 13.53)	0.101
Tobacco Smoking	5.12 (3.49, 7.62)	< 0.001	3.44 (2.23, 5.33)	< 0.001
Calcium Deficiency	6.40 (2.98, 15.91)	< 0.001	3.54 (1.43, 9.87)	0.009
Potassium Deficiency	7.00 (3.29, 17.42)	< 0.001	3.94 (1.63, 10.64)	0.004

Fig. 4.

Odds ratios of top 10 lifestyle exposures identified in clinical notes. Odds ratio of each lifestyle exposure was calculated with adjustment for age and other 9 exposures. * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001.

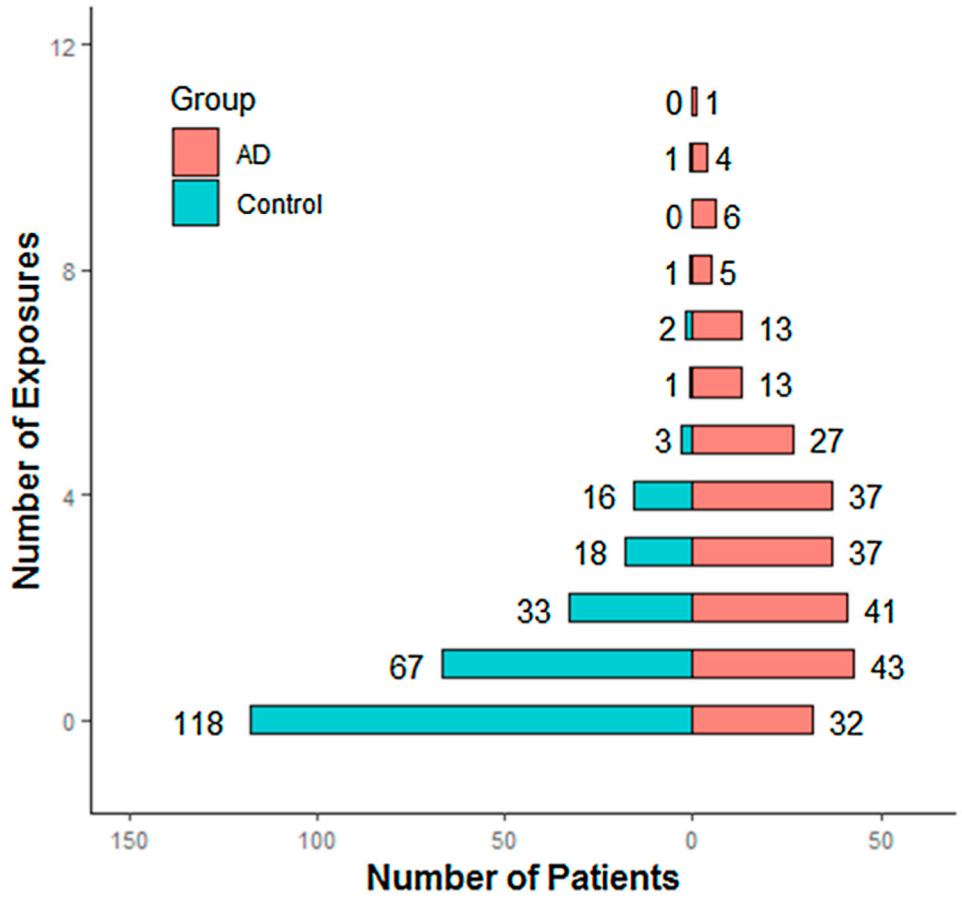


Fig. 5.
Distribution of lifestyle exposures between AD group and control group.

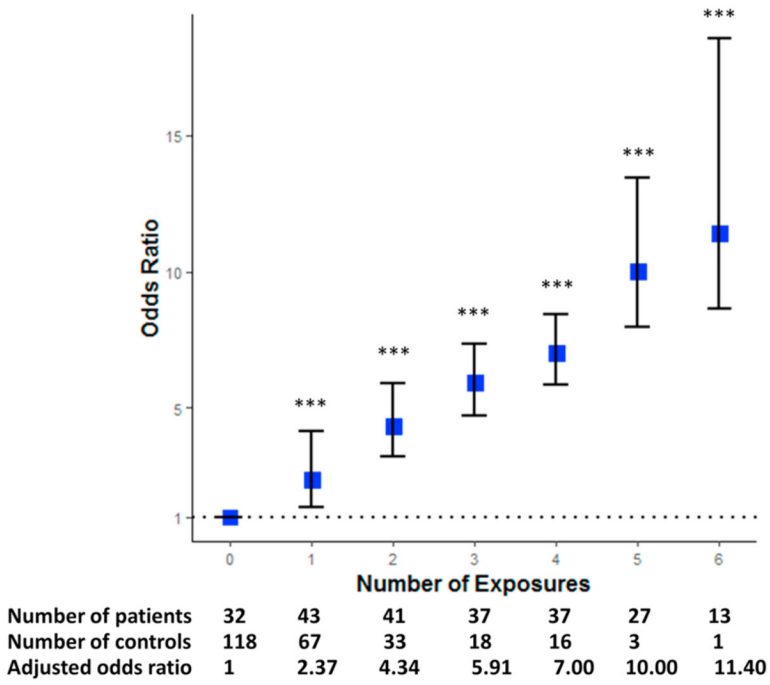


Fig. 6.

Odds of AD according to number of lifestyle exposures. *** p-value < 0.001.

Table 1

Cross-sectional lifestyle studies of Alzheimer's disease.

Study	Study Design	Location	Sample Size	Lifestyle Exposures Investigated	No. of Exposures
Mosconi L, et al (2018)	Cross-sectional	USA	116	Mediterranean diet Physical activity Intellectual activity	3
Berti V, et al, (2015)	Cross-sectional	USA	52	Nutrition pattern	1
Gardener S, et al (2012)	Case-control	Australia	247 cases, 723 controls	Mediterranean diet	1
Gauthier E, et al (2001)	Case-control	Canada	68 cases, 68 controls	Pesticide exposure	1
Rogers MA, et al (1999)	Case-control	USA	23 cases, 23 controls	Aluminum intake	1

Table 2

Demographic Characteristics of the Study Cohort.

Demographics		**Count(%)**		**Statistic**	**P-value**
		AD group (n = 260)	Control group (n = 260)		
Age	< 75	48 (18.4)	69 (26.5)	t = 0.625	0.532
	76 – 85	111 (42.7)	104 (40.0)		
	86 – 95	92 (35.4)	74 (28.5)		
	> 95	9 (3.5)	13 (5.0)		
	Mean (SD)	82.28 (9.47)	81.05 (7.80)		
Sex	Female	127 (48.9)	136	$\chi^2 = 0.492$	0.483
	Male	133 (51.1)	124		
Race	White	253 (97.3)	256 (98.5)	$\chi^2 = 5.174$	0.160
	Black	0 (0.0)	1 (0.3)		
	Asian	4 (1.5)	0 (0.0)		
	Other	3 (1.2)	3 (1.2)		
Marital status	Married	193 (74.2)	173 (66.5)	$\chi^2 = 4.634$	0.201
	Single	8 (3.1)	14 (5.4)		
	Divorced	14 (5.4)	21 (8.1)		
	Widowed	45 (17.3)	52 (20.0)		

Table 3

Lifestyle Exposures Reported in AD Group.

Category	Lifestyle Exposures	No. of records	No. of patients	% of patients
Dietary Factor	High Calorie Diet	1	1	< 1.0
	High Carbohydrate Diet	1	1	< 1.0
	High Meat Diet	6	2	< 1.0
	High Fat Diet	36	3	1.2
	High Pickle Diet	3	2	< 1.0
	Vitamin B Deficiency	1181	28	10.8
	Vitamin D Deficiency	2629	29	11.2
	Vitamin E Deficiency	10	1	< 1.0
	Potassium Deficiency	1454	43	16.5
	Iron Deficiency	1872	25	9.6
	Magnesium Deficiency	382	12	4.6
	Calcium Deficiency	558	38	14.6
	Starvation	22	2	< 1.0
	Dehydration	1932	91	35
	Malnutrition	3567	134	51.5
Daily Activity	Physical Inactivity	6	4	1.5
	Sleep Disorder	25	5	1.9
Substance Abuse	Phencyclidine	24	2	< 1.0
	Tobacco Smoking	1081	145	55.8
	Excess Alcohol	7469	52	20

Table 4

Lifestyle Interventions Strategies in AD Group.

Category	Lifestyle Intervention	All Sections			AD Sections		
		No. of Record	No. of Patient	% of Patient	No. of Record	No. of Patient	% of Patient
Dietary Factor	Low Fat Diet	187	26	10.0			
	Low Salt Diet	379	51	19.6	1	1	< 1.0
	Low Carbohydrate Diet	1	1	< 1.0			
	Low Calorie Diet	1	1	< 1.0			
	Low Cholesterol Diet	200	23	8.8			
	Vitamin B Supplement	4	2	< 1.0			
	Vitamin C Supplement	4	1	< 1.0			
	Vitamin D Supplement	226	21	8.1			
	Vitamin E Supplement	1163	4	1.5			
	Potassium Supplement	62	15	5.8			
	Iron Supplement	36	10	3.8			
	Zinc Supplement	4	1	< 1.0			
	Magnesium Supplement	7	2	< 1.0			
	Calcium Supplement	381	38	14.6			
	Dietary Supplement	1843	148	56.9	33	4	1.5
	Nutrition Supplement	78	16	6.2			
	Docosahexaenoic Acid	303	6	2.3			
	Vegetables	3141	89	34.2			
	Fatty Fish/Fish Oil	16528	94	36.2	1	1	< 1.0
Daily Activity	Physical Activity	18579	227	87.3	123	23	8.8
	Cognitive Activity	238	46	17.7	45	7	2.7
	Benzodiazepines	2672	17	6.5	19	1	< 1.0
Substance Abuse	Smoking Cessation	1200	137	52.7			

Table 5

Lifestyle Exposures Reported in Control Group.

Category	Lifestyle Exposures	No. of Records	No. of Patients	% of Patients
Dietary Factor	High Iron Diet	1	1	< 1.0
	Vitamin B Deficiency	325	5	1.9
	Vitamin D Deficiency	4508	11	4.2
	Vitamin K Deficiency	10	1	< 1.0
	Potassium Deficiency	329	7	2.7
	Iron Deficiency	581	15	5.8
	Magnesium Deficiency	68	4	1.5
	Calcium Deficiency	96	7	2.7
	Dehydration	379	35	13.5
	Malnutrition	713	49	18.8
Daily Activity	Sleep Disorder	6	2	< 1.0
Substance Abuse	Tobacco Smoking	159	55	21.2
	Excess Alcohol	2884	25	9.6

Table 6

Lifestyle Exposures not found in Clinical Notes.

Category	Without UMLS concepts	With UMLS concepts
Dietary Factor	Diabetogenic diet	High copper diet
	High advanced glycation end products diet	High zinc diet
	High arachidonic acid	High methionine diet
	High unfermented soy	Low tryptophan diet
	Linoleic acid deficiency	Glucose deprivation
	Early life nutrient restriction	Selenium deficiency
	Glutathione depletion	Industrialized/preserved food
	Low cocoa	Monosodium glutamate
	Low coffee	Cysteine
	Low flavonols	
	Low fruit	
	Menadione	
	Diacetyl	
Substance Abuse	Amphetamine	
	MDMA	
	Cocaine/opiates	