Table of content

1. Welcome

Dear student,

We are happy to welcome you to Aletta's Regional Year Challenge program! This toolkit will provide you with all the information you need regarding the challenge: *active and healthy lifestyle*.

We hope you will have a great time during the program and come up with just as great solutions.

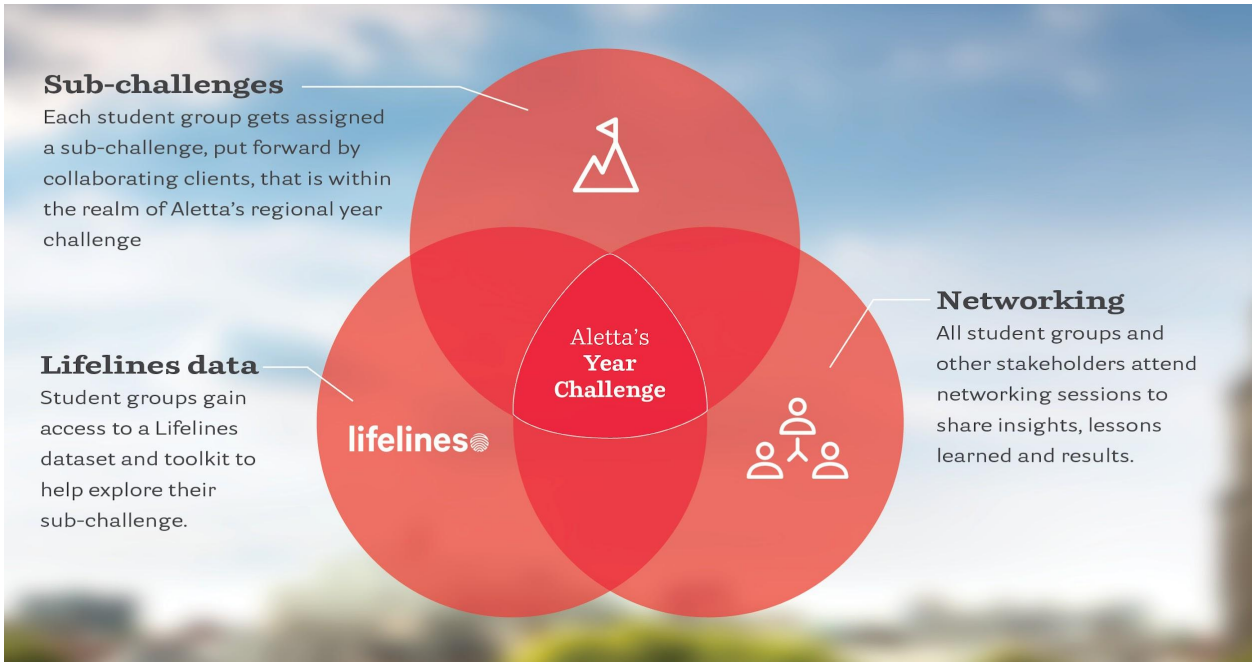
Good luck!

Aletta School for Public Health
Lifelines
Innovation Center UMCG

2. Challenge & Context

This year a new edition of Aletta's Regional Year Challenge will start. As a student, you are invited to participate in this program and work on this year's challenge: '*active and healthy lifestyle'*. During the program you are invited to join our events, collaborate with other students, work together with (public) health stakeholders in the area and work on your solution for our regional challenge.

You will have access to a Lifelines public health dataset to explore and deepen your understanding of the challenge. Through networking events all students, clients and end-users will be able to share insights and results. Results may include visualisations of data, policy advice and practical innovative on-the-ground solutions, such as suggestions for redesigning services.



3. Data

For this challenge you can use data from the Lifelines cohort. Lifelines is a large, multi-generational, prospective cohort study that includes over 167,000 participants (10%) from the northern population of the Netherlands. Within this cohort study the participants are followed over a 30-year period. Every five years, participants visit one of the Lifelines sites in the northern parts of the Netherlands for an assessment. During these visits, several physical measurements are taken and different biomaterials are collected. As part of the assessment, participants are asked to fill out comprehensive questionnaires. In between assessments, participants are invited to complete follow-up questionnaires approximately once every 1.5 years.

Lifelines offers multiple data sources to support you in this challenge. You can read more about these sources within this chapter.

Lifelines interactive map
The Lifelines interactive map is a visualization of data collected from the first five years. The map is easily accessible and gives you a quick glimpse in how the data differs between the various geographical regions. Have a look at for example the spread of *BMI* and look at data regarding *sports* and *stress*. While you can not connect two different variables, they do give you an idea of areas where you can work on health potential.

Visualizations also help to get insights in peculiar data, try looking up the data on *teeth*, what do you make of that?!

https://inzicht.lifelines.nl/
username: Lifelines, password: Maaktonderzoekmogelijk

Lifelines corona barometer
The Lifelines corona barometer is a visualization which is being updated with all of the Covid-19 related questions that were collected throughout 2020 and 2021. In total over 20 questionnaires were sent out to participants and this enormous data collection resulted in insights which were used for scientific studies but also governmental policies. The corona barometer allows more in-depth filters and comparisons on socio-demographic variables, such as age, gender and education.

https://coronabarometer.nl/

Researcher publications overview
Interested in learning if the challenge you want to solve has a sound scientific basis? Head over to our publications overview and see if you can find research results that match your topic.

https://www.lifelines.nl/researcher/publications

Lifelines wiki
The Lifelines wiki can provide you with more information about the Lifelines cohort study design and metadata descriptions of the available data collection.

http://wiki.lifelines.nl

Lifelines public health dataset

The Lifelines public health dataset is the data you have at your disposal to work on your challenge. The data is aggregated on zip codes and on age groups in order to preserve the privacy of participants. The dataset contains a variety of topics on public health which are explained next.

Data availability (complete versus continuous)

The data is offered in two variants: complete and continuous. The complete variant contains all respondents that participated at one or both time points (T1 or T2), while the continuous variant only contains data of participants who have participated at both time points (T1 and T2). This means that the continuous variant only has T1 data of participants who also had data at T2.

The continuous variant is useful if you want to compare data between the two timepoints and you want to be certain that you're looking at the same group of participants. If you are not interested in change over time for a single group of participants, you can use the complete variant, which contains all of the respondents at both T1 and T2. This includes participants that have data at T2, but may not have data at T1.

Default variables

To be able to distinguish the data and create groups of participants to find possible differences, you will have access to basic demographic data such as zip codes, age and gender.

Variable	**Label**
ZIPCODE*	Postal code 4 Baseline visit 1
AGE**	Current age
GENDER	Gender (1=male, 2=female)
GROUP_SIZE_CAT	Group size in categories (0-200=1, 200-300=2, 300-400=3 ... 1300-1400=13, 1400-1500=14, 1500+=15)
AGE_CAT*	Age in categories (15-20=1, 20-25=2, 25-30=3 ... 85-90 = 15, 90-95=16, 95+=17)
GENDER_MALE*	Male participants in percentage
GENDER_FEMALE*	Female participants in percentage
AGE_T1	Age at T1
AGE_T2	Age at T2

* Available in dataset(s) aggregated on zip code, ** Available in dataset aggregated on age

Anthropometry

Anthropometry (section: physical state) was measured in all Lifelines participants aged 8 years and older during assessments:

- 1A Visit 1 (n = ~163.000)
- 2A Visit 1 (n = ~119.000)

Height, weight, waist and hip circumference were assessed. Participants were also asked to self-report their height and weight.

Variable	**Label**
BMI_T1	Body Mass Index (kg/M^2) at baseline as mean (T1)
WEIGHT_T1	Weight in kg at second assessment as mean (T1)
HIP_T1	Hip circumference in cm at baseline as mean (T1)
HEIGHT_T1	Height in cm at baseline as mean (T1)
WAIST_T1	Waist circumference in cm at baseline as mean T1)

BMI_T2	Body Mass Index (kg/M^2) at second assessment as mean (T2)
WEIGHT_T2	Weight in kg at second assessment as mean (T2)
HIP_T2	Hip circumference in cm at second assessment as mean (T2)
HEIGHT_T2	Height in cm at second assessment as mean (T2)
WAIST_T2	Waist circumference in cm at second assessment as mean (T2)

Education

Education (section: demographics) is collected over various time points and has been coded into educational attainment. For the public health dataset we show the percentage of participants within their group of aggregation (on zip code or age) whose highest education is 'lower education'. Lower education includes:

- No education (did not finish primary school)
- Primary education (primary school, special needs primary school)
- Lower or preparatory secondary vocational education (such as lts, leao, lhno, vmbo)
- Junior general secondary education (such as mavo, (m)ulo, mbo-short, vmbo-t)

Variable	**Label**
EDUCATION_LOWER_T1	Education at baseline in percentage where the highest obtained degree is lower education (ranging from no education to junior general secondary education) (T1)
EDUCATION_LOWER_T2	Education at second assessment in percentage where the highest obtained degree is lower education (ranging from no education to junior general secondary education) (T2)

Work

Data on work (section: demographics) is collected over various time points and shows the percentage of participants who have paid work for at least a few hours a week. In the wiki you will find more information on this data point by searching for variable WORK2A2, which is part of a series of questions on employment status.

Variable	**Label**
WORK_T1	Paid work (for at least a few hours a week) at baseline in percentage (T1)
WORK_T2	Paid work (for at least a few hours a week) at second assessment in percentage (T2)

Quality of life

Quality of life (section: wellbeing and subjective health) is measured as a variable from the standardized questionnaire 'Short Form Health Survey'. This is a popular question for research (especially to measure over time) and is used in a variety of questionnaires, including additional studies. For the public health dataset the percentage of participants that have reported to have a low quality of life (poor/mediocre) is given.

Variable	**Label**
LOW_QUALITY_OF_LIFE_T1	Low quality of life (poor/mediocre) at baseline in percentage (T1)
LOW_QUALITY_OF_LIFE_T2	Low quality of life (poor/mediocre) at second assessment in percentage (T2)

Blood pressure and hypertension

Data on hypertension (section: physical state) is collected by a triangulation of self-report health questionnaires, medication by prescription questionnaires, which are handed in by the participants during the second Lifelines visit, and blood pressure measures. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were recorded while the participant was in supine resting position. With an automatic blood pressure monitor (DinaMap, PRO 100V2) assessments were taken every minute during 10 minutes. The last three measures were used.

Variable	**Label**
DBP_T1	Diastolic Blood Pressure in mm hg at baseline as mean (T1)
DBP_T2	Diastolic Blood Pressure in mm hg at second assessment as mean (T2)
HBF_T1	Pulse rate in beats per minute at baseline as mean (T1)
HBF_T2	Pulse rate in beats per minute at second assessment as mean (T2)
MAP_T1	Mean Average Pressure at baseline as mean (T1)
MAP_T2	Mean Average Pressure at second assessment as mean (T2)
SBP_T1	Systolic Blood Pressure in mm hg at baseline as mean (T1)
SBP_T2	Systolic Blood Pressure in mm hg at second assessment as mean (T2)
HTN_MED_T1	Hypertension medication at baseline first visit in percentage (T1)

Biosamples: blood (glucose and cholesterol)

The following variables were assessed in the plasma of fresh blood samples collected from Lifelines participants during 1A Visit 2 and 2A Visit 2 (section: physical state).

Variable	**Label**

CHO_T1	Cholesterol (mmol/L) assessed on samples from baseline as mean (T1)
GLU_T1	Glucose (mmol/L) assessed on samples from baseline as mean (T1)
CHO_T2	Cholesterol (mmol/L) assessed on samples from second assessment second visit as mean (T2)
GLU_T2	Glucose (mmol/L) assessed on samples from second assessment as mean (T2)

Health and lifestyle

Lifelines Diet Score

Dietary intake of Lifelines participants aged 13 years and older was assessed with a food frequency questionnaire (FFQ) developed for Lifelines by the department of Human Nutrition and Health at WUR (section: nutrition). Since this resulted in a long and time-consuming questionnaire, the questionnaire was divided into four parts and called the Flower Food Frequency Questionnaire. For more information about the rationale for the flower FFQ see Brouwer-Brolsma et al. (2017). The Lifelines Diet Score is a derivative score indicating the quality of one's diet.

Alcohol intake

Alcohol intake is a calculation based on various items from the food frequency questionnaire (FFQ) (section: nutrition) which represents the total alcohol intake per day in grams.

Energy (calory) intake

Energy intake is a calculation based on various items from the food frequency questionnaire (FFQ) (section: nutrition) which represents the total energy intake per day in grams.

Sports

Sports (section: lifestyle and environment) indicates how many participants participate in sports at least once a week.

Cycling to work

Cycling to work is a calculation based on various items from the SQUASH instrument (see also Physical Activity) (section:). For this dataset, the variable indicates how many participants cycle to work at least once a week.

Volunteering

Organized volunteer work (section: lifestyle and environment) indicates how many participants participate in organized volunteer work related to a volunteering organization or association

Pregnancy

Pregnancy (section: reproduction and development) is measured as an average of how many participants have given birth (within the aggregated group)

Sleep quality

Data on sleep quality (section: lifestyle and environment) has been reviewed by researchers from the Erasmus MC who have developed a PSQI derivative for the Lifelines cohort. The variable included in the public health dataset is the percentage of participants who indicated to experience 'poor sleep quality'.

Physical Activity

Based on the raw data from the SQUASH instrument in 1A Questionnaire 2, the sum scores for physical activity per activity type and overall were calculated for all adult participants who sufficiently filled in the questionnaire (sections: lifestyle & environment and secondary & linked variables).

Variable	**Label**
PACKYEARS	Cumulative smoking history: 1 packyear = 20 cigarettes per day for 1 year (or 10 cigarettes for 2 years, or 1 cigarette for 20 years). NB. Cigars are regarded as 3 cigarettes. (T1)
SMOKING	Currently smoking at baseline in percentage (T1)
LLDS	Lifelines Diet Score as continuous variable, with a range between 0 and 48, at baseline as mean (T1)
Quintile_LLDS	Lifelines Diet Score as categorical variable, with Q1 referring to poorest and Q5 referring to highest diet quality as mean (T1)
ALCOHOL_INTAKE_T1	Total alcohol intake in grams per day as a mean at baseline (T1)
KCAL_INTAKE_T1	Total energy intake per day expressed in kcal as a mean at baseline (T1)
MWK_VAL	Minutes of weekly physical activity on moderate and vigorous intensity level at baseline as mean (T1)
SCOR_VAL	Score for weekly physical activity on moderate and vigorous intensity level, based on the sum of minutes per activity times the intensity of the specific activity at baseline as mean (T1)
MWK_NO_VAL	Minutes of weekly physical activity on moderate and vigorous intensity level, in leisure time and commuting domains (but not occupational) at baseline as mean (T1)

SCOR_NO_VAL	Score for weekly physical activity on moderate and vigorous intensity level, in leisure time and commuting domains (but not occupational), based on the sum of minutes per activity times the intensity of the specific activity at baseline as mean (T1)
PREGNANCIES	Number of times participant has given birth at baseline as mean (T1)
SLEEP_QUALITY	Poor sleep quality as measured through the PSQI at baseline in percentage (T1)
SPORTS_T1	Participates in sports at baseline in percentage (T1)
CYCLE_COMMUTE_T1	Cycles to work (commute) at least 1 day during the week at baseline in percentage (T1)
VOLUNTEER_T1	Participates in organized volunteer work at baseline in percentage (T1)

Conditions, illnesses and diseases

Respiratory disease (such as asthma & COPD)

Lifelines participants were asked whether they had and/or were treated for various respiratory diseases (section: diseases & symptoms). Asthma and COPD have been defined by looking at self-reported phenotypes, medication intake and spirometry during visits.

Metabolic disorder (such as type 2 diabetes)

Data on metabolic disorders (section: diseases & symptoms) were collected by questionnaire at baseline, ~1.5 years after baseline, ~3.0 years after baseline and at the second assessment (~4.5 years after baseline). Additionally, for a more accurate disease classification, laboratory measurements collected at baseline and at the second assessment were used.

Osteoarthritis

Osteoarthritis, or joint degradation, is measured as the percentage of participants that suffer from this condition (section: diseases & symptoms).

Chronic Fatigue Syndrome (CFS)

Identify subjects who fulfill the diagnosis of Chronic Fatigue Syndrome (CFS) based on 1994 Centers for Disease Control and Prevention (CDC) criteria (section: diseases & symptoms). To meet the criteria, the following 3 keys are required: i) experienced chronic fatigue for 6 or more months, ii) report that the fatigue significantly interfered with daily activities and work, and iii) report that concurrently four or more of the eight additional symptoms are present (sore throat, tender lymph nodes, muscle pain, joint pain, headaches, unrefreshing sleep, unusual fatigue after exertion, forgetfulness and/or difficulty concentrating).

Fibromyalgia

The relevant symptoms for the ACR criteria are collected only at the second assessment (2A). It requires both second assessment questionnaire 2A1 and 2A2 for adults (age 18 and older).

The Checklist Individual Strength (CIS) is a 20-item fatigue questionnaire developed in Dutch (section: diseases & symptoms). The questionnaire has been translated into multiple languages and is used to set diagnostic criteria for various illnesses, including chronic fatigue syndrome/myalgic encephalomyelitis (ME/CFS).

The Widespread Pain Index quantifies the extent of bodily pain. The WPI is required to identify individuals that meet the diagnostic criteria of fibromyalgia in epidemiological studies (section: Diseases & symptoms). Symptoms experienced by Lifelines participants were assessed using the Symptom Checklist (SCL-90), a multidimensional self-report instrument (section: Diseases & symptoms) that was found to be suitable for use in large-scale studies including elderly people.

Irritable Bowel Syndrome (IBS)

Identify subjects who fulfill the diagnosis of Irritable Bowel Syndrome (IBS) based on ROME III criteria. Lifelines repeatedly assessed the presence of IBS by self-report. The Rome III IBS Diagnostic Questionnaire is a survey to assess whether individuals meet the diagnostic criteria for Irritable Bowel Syndrome (IBS) (section: diseases & symptoms). The relevant symptoms for the ROMEIII criteria are collected only at the second assessment (2A) for adults (age 18 and older).

Variable	**Label**
RESPIRATORY_DISEASE_T1	Respiratory disease at baseline in percentage (T1)
METABOLIC_DISORDER_T1	Metabolic disorder at baseline in percentage (T1)
METABOLIC_DISORDER_T2	Metabolic disorder at second assessment in percentage (T2)
OSTEOARTHRITIS_T1	Joint degradation (osteoarthritis) at baseline in percentage (T1)
DIAG_CFS_CDC	Fulfill CDC Chronic Fatigue Syndrome criteria in percentage (T2)
DIAG_FIBROMYALGIA_ACR	Fulfill ACR Fibromyalgia criteria in percentage (T2)
DIAG_IBS_ROME3	Fulfill ROMEIII Irritable bowel syndrome criteria in percentage (T2)

Personality

Lifelines implemented two abbreviated versions of the NEO at baseline, focusing on the domains of neuroticism, extraversion and conscientiousness. The NEO-PI-R is a self-report instrument that measures the five most important domains of personality: neuroticism, extraversion, openness, agreeableness and conscientiousness (section: mental health).

Variable	Label
C_SUM_T1	Sum of all items within facet Competence. Complete cases only, as mean (T1)
A_SUM_T1	Sum of all items within facet Anger-hostility. Complete cases only, as mean (T1)
SC_SUM_T1	Sum of all items within facet Self-consciousness. Complete cases only, as mean (T1)
I_SUM_T1	Sum of all items within facet Impulsivity. Complete cases only, as mean (T1)
E_SUM_T1	Sum of all items within facet Extraversion. Complete cases only, as mean (T1)
SD_SUM_T1	Sum of all items within facet Self-discipline. Complete cases only, as mean (T1)
V_SUM_T1	Sum of all items within facet Vulnerability. Complete cases only, as mean (T1)
D_SUM_T1	Sum of all items within facet Deliberation. Complete cases only, as mean (T1)

Stress

Data on past-year stress exposure (section: mental health) at baseline and follow-up (3 time points: follow-up questionnaire 1 and 2 and questionnaire at second assessment).

Two stress exposure questionnaires were administered: the List of Threatening Experiences (LTE) measuring acute stressful life events (or SLE), and the Long-term Difficulties Inventory (LDI) measuring long-term difficulties (LTD) i.e. more chronic stressors. The questionnaires were part of the questionnaire packs to be administered at home. Both concern stressors in the past year. The validation and the reliability of the LDI was performed in a study by Rosmalen et al.

The LTE lists 12 stressful life events (e.g. death of a relative, serious disease) and asks whether participants experienced such an event in the past year (no=2, yes=1). Items 13 (“did you experience any other major life events in the past year?”) and 13A (“can you briefly describe this event?”) are discarded.

The LDI lists 12 potential sources of chronic stress (e.g. financial difficulties, work-related stress, strained relationships) and asks how much the participant was affected by this type of stress (not=1, somewhat=2, much=3).

Variable	Label
LTE_SUM_T1	Total number of stressful life events that happened to participant in the past year at baseline as mean (T1)
LDI_SUM_T1	Total amount of stress from long-term/chronic stressors experienced by the participant at baseline as mean (T1)
LTE_SUM_T2	Total number of stressful life events that happened to participant in the past year at second assessment as mean (T2)

LDI_SUM_T2	Total amount of stress from long-term/chronic stressors experienced by the participant at second assessment as mean (T2)

Neighbourhood socio-economic status

Lifelines collected demographic variables that describe the basic socio-economic situation of participants. Lifelines has generated a list of municipal codes for each participant, one code for each known home address starting in 2006 (sections: demographics and secondary & linked variables). These municipal codes correspond to the codes used by Statistics Netherlands (CBS) to collect data ("kerncijfers") about Dutch neighbourhoods, organised in tables per year. Lifelines provides both the municipal code per consecutive home address of each participant (since 2006), and the CBS tables with the corresponding neighbourhood data for the years 2008-2019.

Variable	**Label**
NSES_YEAR	Year for which the NSES score was calculated as mean
NSES	Neighborhood socio-economic status score according to CBS Statistics Netherlands, based on inhabitants' educational level, income and job prospective as mean

Anxiety and depression (MINI)

The Mini International Neuropsychiatric Interview (MINI) is a brief, reliable and valid structured diagnostic interview for diagnosing psychiatric disorders, compatible with international diagnostic criteria such as the DSM-IV and ICD-10.

The MINI (section: mental health) was performed in adult and elderly Lifelines participants as a face-to-face interview with a nurse practitioner on location during 1A Visit 1, and as a digital questionnaire on location during 2A Visit 1.

Variable	**Label**
MENTAL_DISORDER_T1	Cumulative mean diagnosis of major depressive disorder, dysthymia, generalized anxiety disorder, panic disorder, agoraphobia and social anxiety disorder / social phobia at baseline in percentage (T1)
MENTAL_DISORDER_T2	Cumulative mean diagnosis of major depressive disorder, dysthymia, generalized anxiety disorder, panic disorder, agoraphobia and social anxiety disorder / social phobia at second assessment in percentage (T2)

External data

Using the demographic groups within the Lifelines dataset or the zip codes you can also pair other external data for your sub challenge. For example, if you want to use a different geographical layout you can pair the Dutch neighbourhood codes to your dataset using the zip

codes. You can find this data via Statistics Netherlands: https://www.cbs.nl/nl-nl/maatwerk/2020/39/buurt-wijk-en-gemeente-2020-voor-postcode-huisnummer. On the website of the Dutch government you will also be able to find other open datasets: https://data.overheid.nl/.

4. Example cases

As an example, let's take a sub challenge that could have been defined within our year challenge: *'How can we promote and improve physical activity among our citizens?'*. A province or municipality could be interested in learning how they can engage their citizens in exercising more to promote a healthy lifestyle.

Step 1: Orientation
To get some orientation on the subject, we could first look at some variables in the interactive map and see if there's anything related to physical activity. Three possible questions stand-out:

- 'Sport': How many hours a week do you exercise on average? (figure 1)
- 'Limitations': Are you currently being limited in bending or kneeling? (figure 2)
- 'Physiotherapist': Have you been in touch with a physiotherapist the past 12 months? (figure 3)

If we look at the visualization of the data in figures 1 to 3 we can see that, on average, citizens in the eastern part of the province of Groningen and Drenthe exercise less than other Lifelines participants. Similar to the data on limitations in bending and kneeling, we can see that the same regions in Groningen and Drenthe report a higher percentage than in other regions. This might indicate that such data is related. To determine if this group would also go to a physiotherapist more often (because of their limitations), we can look at recent visits (in the past 12 months). This however does not show such a strong contrast (which can have multiple reasons, for example socioeconomic status, healthcare plans, or not feeling the need to seek support).

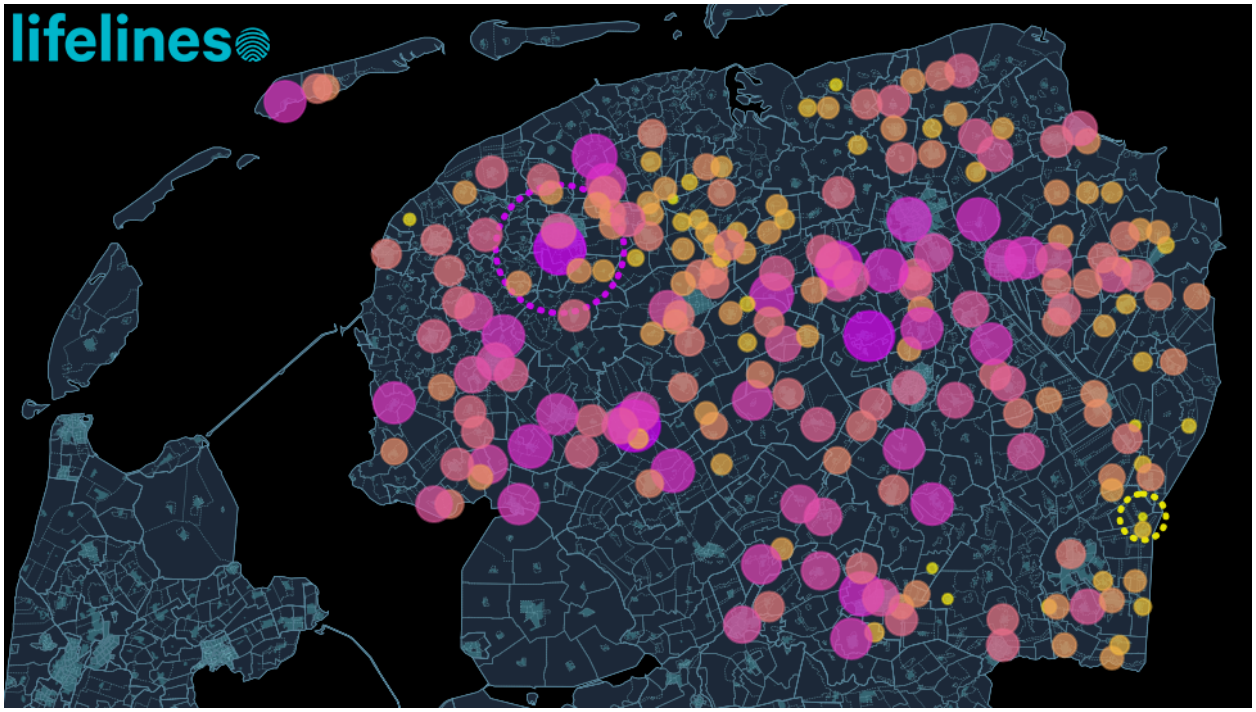


Figure 1: ‘Sport’: How many hours a week do you exercise on average?

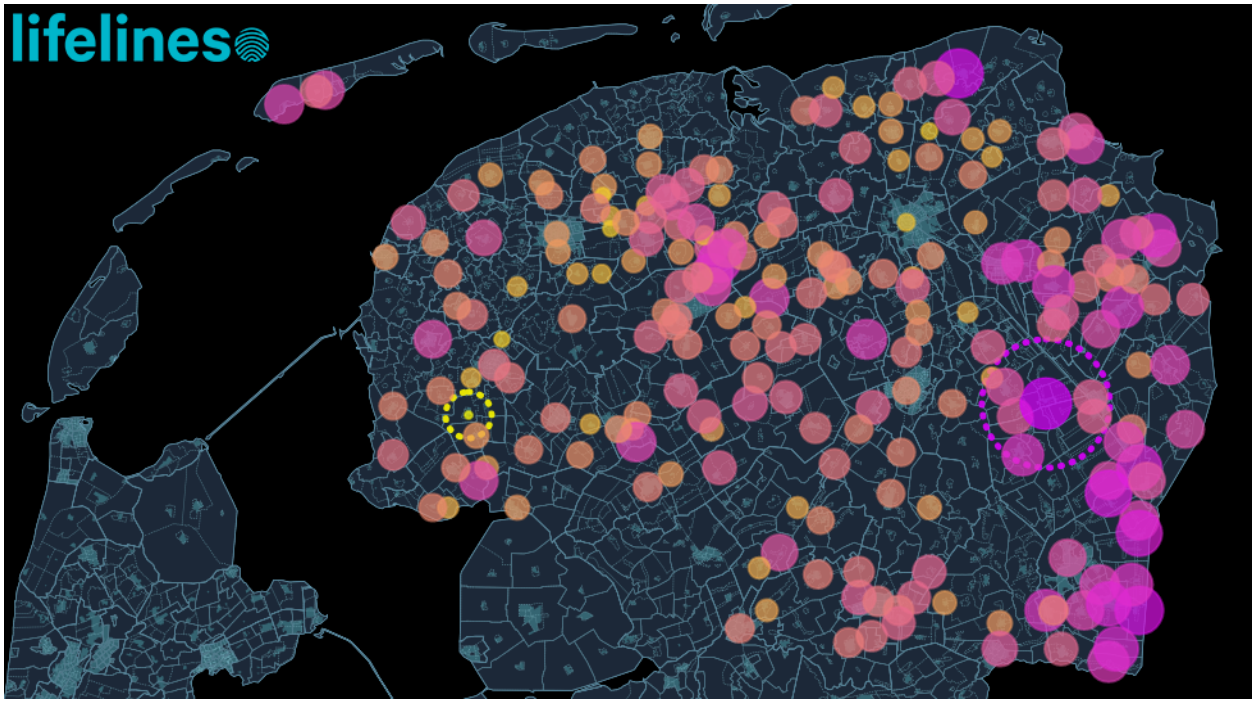


Figure 2: ‘Limitations’: Are you currently being limited in bending or kneeling?

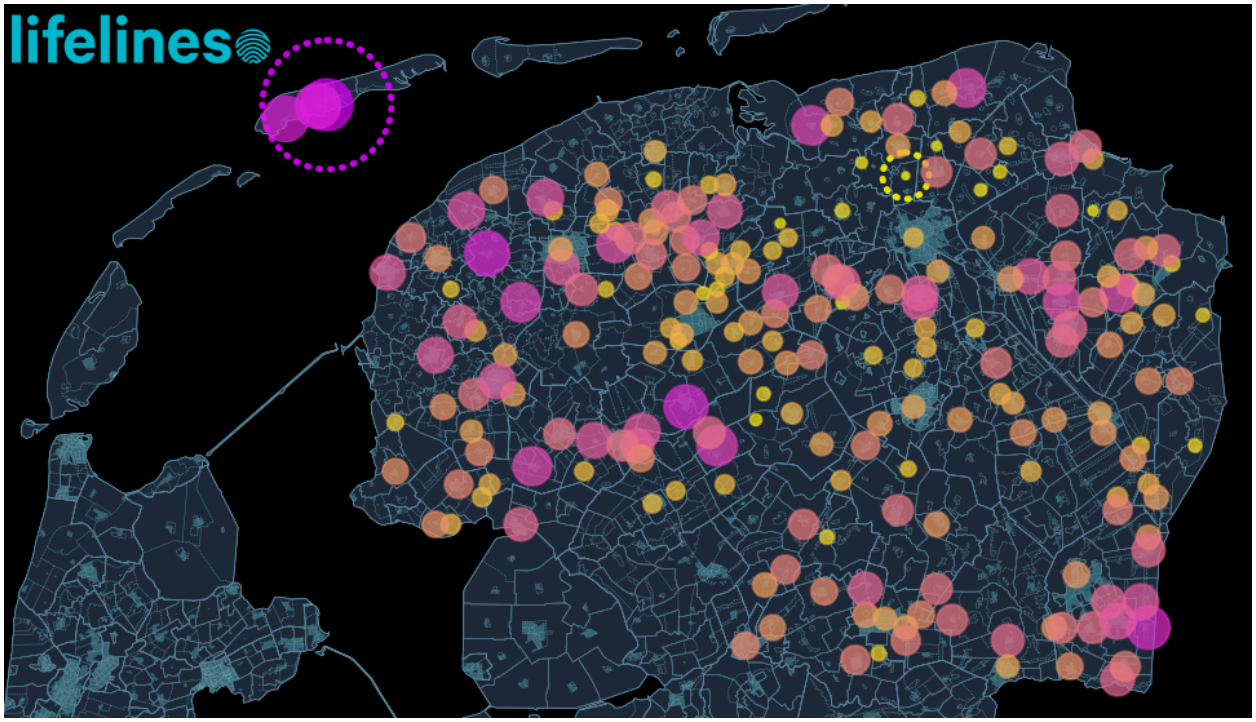


Figure 3: 'Physiotherapist': Have you been in touch with a physiotherapist the past 12 months?

Step 2: Context

Now that there might be a lead and a possible difference between two groups within the same dataset, we need to learn how such data can be interpreted. This means we need to get a better understanding of what physical activity means and how it is measured and analysed (within Lifelines). A first good step would be to find out what other researchers have examined. To do so, we can go to the publication overview and look for 'physical activity' (figure 4, figure 5).

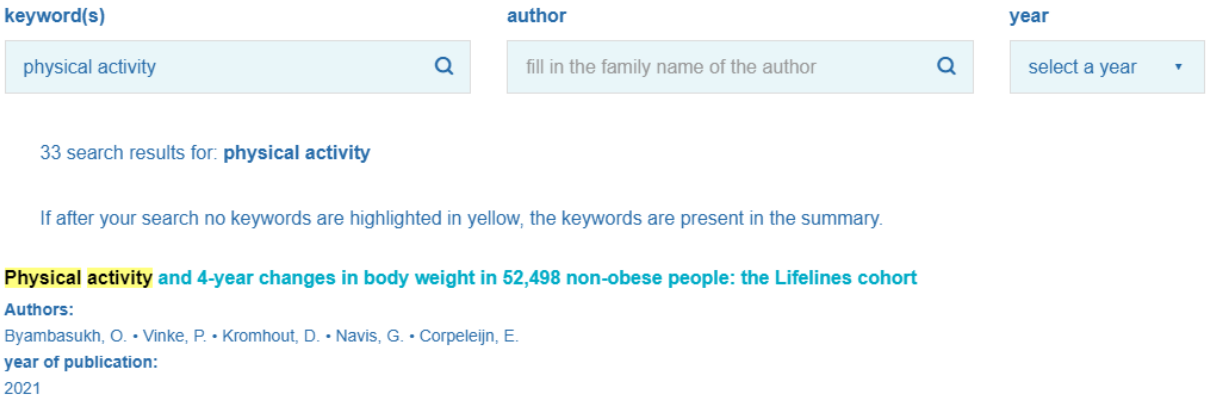


Figure 4: Publication overview on 'physical activity' showing 33 results.

Implementing individually tailored prescription of physical activity in routine clinical care: protocol of the Physicians Implement Exercise = Medicine (PIE=M) development and implementation project

ABSTRACT:
Background:
The prescription of physical activity (PA) in clinical care has been advocated worldwide. This 'exercise is medicine' (E=M) can be used to prevent, manage and cure various lifestyle-related chronic diseases. Due to several challenges, E=M is not yet routinely implemented in clinical care.

Figure 5: One publication leads to another program (PIE=M), which can be a new source of information on actual (practical) implementation/measures

Looking through the publications will help to understand what data has been used and how it has been analysed. This could be data that is also included in the public health dataset. To get more details on how the data has been collected we can visit the Lifelines wiki and search for 'Physical activity'. This will bring us to a page on how the data has been collected and how it is calculated (figure 6). This is important to know, because 'exercising' can be subjective and we will need to have a good definition before we can report on results. For example, can you find if gardening counts as exercise?

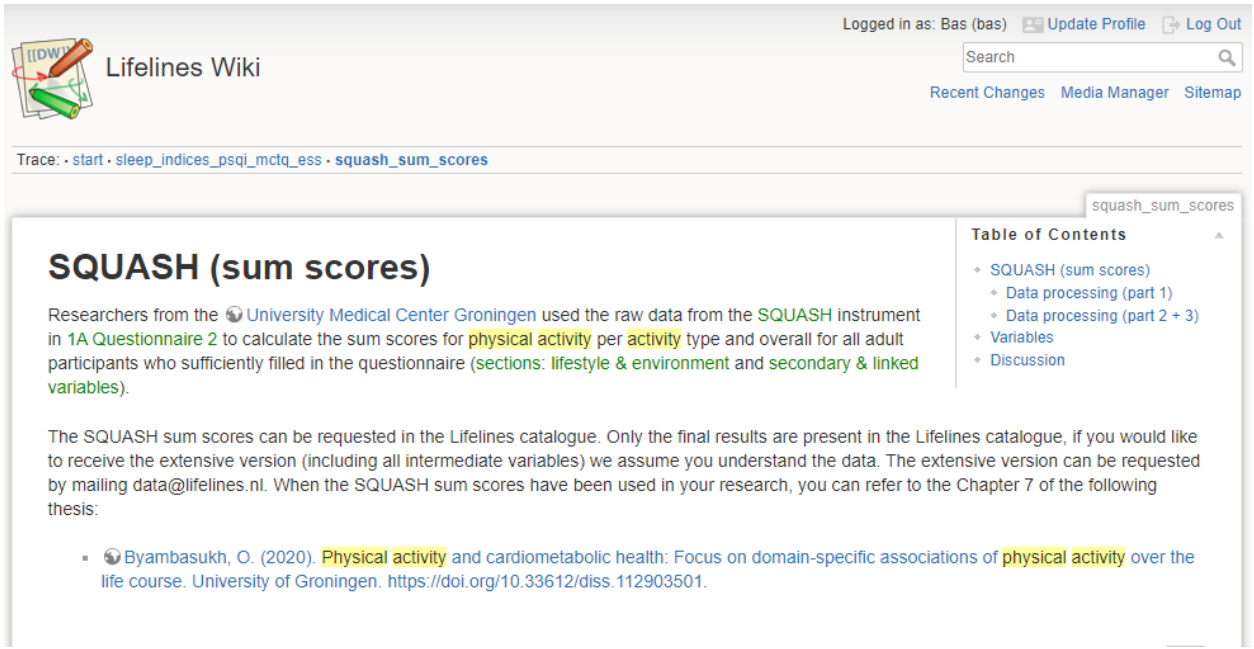


Figure 6: dive into the details on variables and calculations through the wiki

Step 3: Data

Now that we have a direction on what to look for and have an understanding of how the data has been collected and used, we can look at the data ourselves. Looking for data that corresponds to physical activity and what we've found, we might be able to find more answers when looking at the following three variables:

Variable	Label
MWK_VAL	Minutes of weekly physical activity on moderate and vigorous intensity level at baseline as mean (T1)
SCOR_VAL	Score for weekly physical activity on moderate and vigorous intensity level, based on the sum of minutes per activity times the intensity of the specific activity at baseline as mean (T1)
OSTEOARTHRITIS_T1	Joint degradation (osteoarthritis) at baseline in percentage (T1)

To start, it can be helpful to create a graph to correlate variables that may have a relationship. For example, joint degradation is most likely related to age, which we can confirm in the dataset with the plot in figure 7:

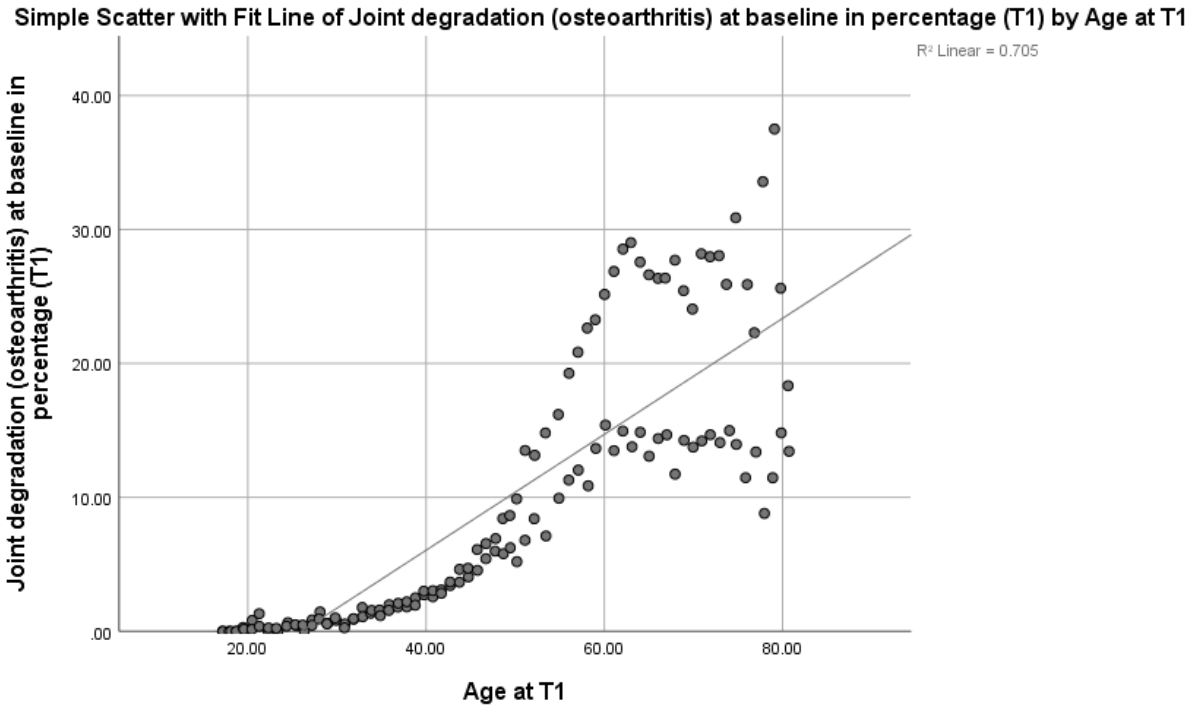


Figure 7: surfacing relationships through graphs; Age on the X-axis and Joint degradation on the Y-axis

Interestingly, from age 60 we can see two groups forming. Let's add gender to see if it is reported differently among men and women (figure 8).

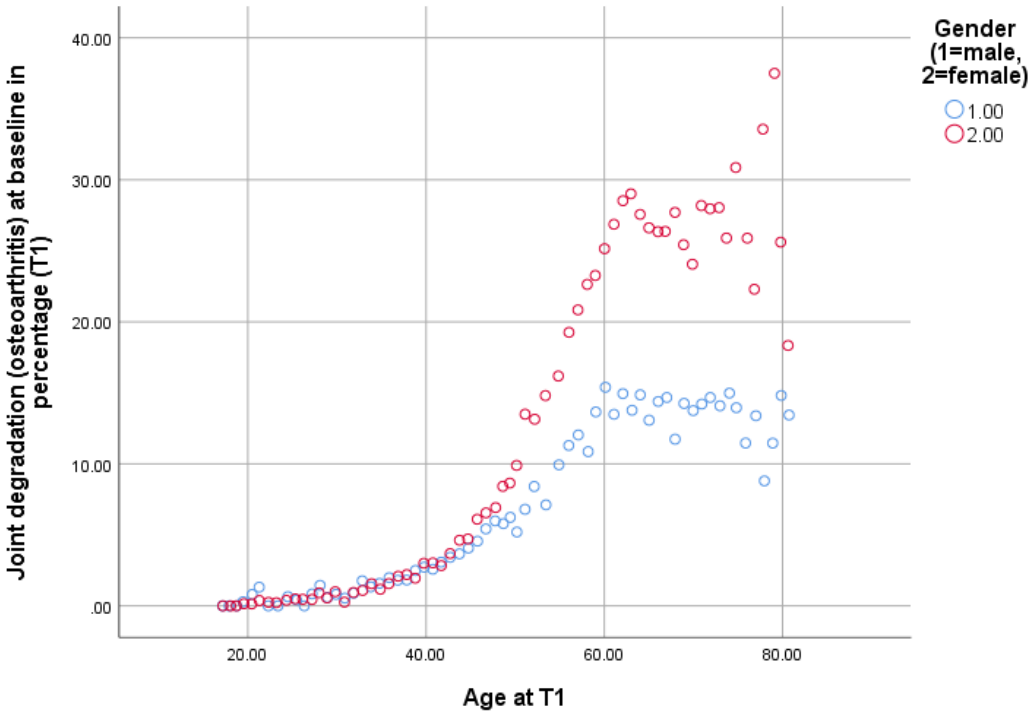


Figure 8: Joint degradation per age with gender distinction

Apparently women report more osteoarthritis than men after turning 60, which could indicate that the disease is more prevalent for older women or that there may be a bias in the self-reported data.

A similar analysis can be done for age and physical activity, of which we can see the results in figure 9.

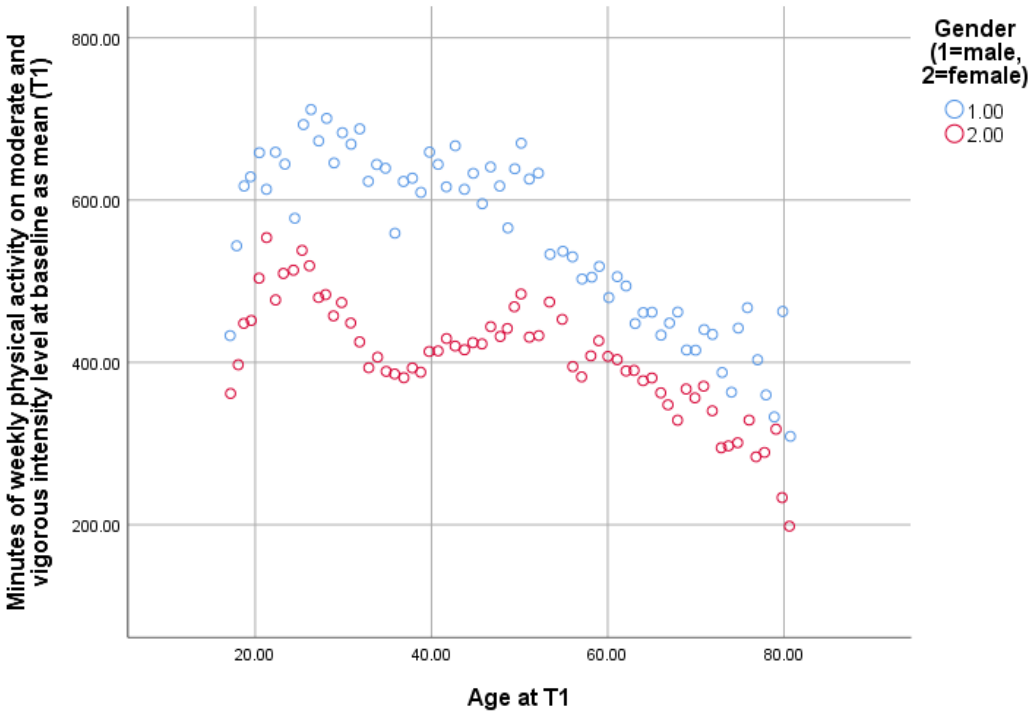


Figure 9: Physical activity per age

Interestingly enough we see a similar gender distinction in physical activity. It also shows us that, as can be expected, physical activity declines as we grow older which is also where the gap between men and women becomes less distinguishable. These results may show that men perform more physical work through their occupation than women, resulting in a higher level of physical activity.

Based on our first figures we saw that the eastern part of Groningen and Drenthe could benefit the most from promoting physical activity. To rule out if this is also age related, we can use the dataset containing zip codes and calculate the same scatter plot for only this region. A simplified version of doing so is filtering the zip codes on the first two digits based on this map: https://postcodebijadres.nl/postcodes-nederland. This map shows us that zip codes starting with 95 and 96 contain the eastern part of Groningen and a small part of Drenthe. By creating a new variable based on these first two numbers we can quickly create a new plot as seen in figure 10.

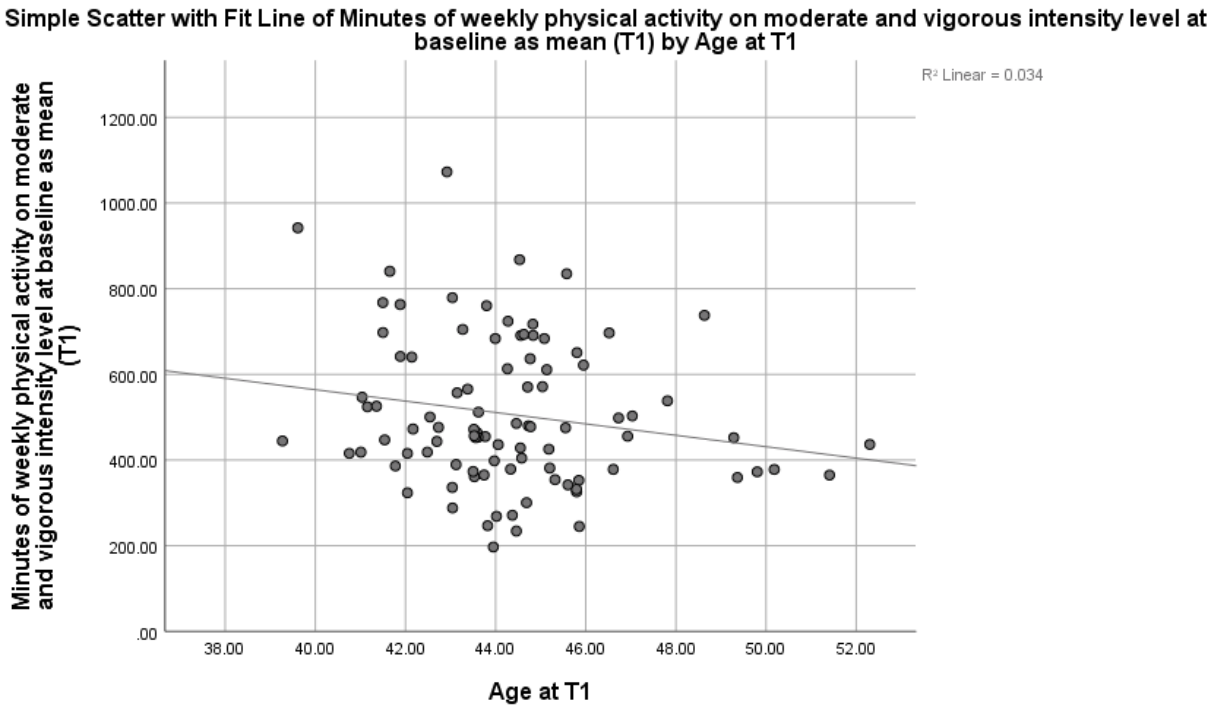


Figure 10: Physical activity per age for zipcodes 95* and 96*

The downward trend that we saw in Figure 9 (activity declines per age) is still visible, but it is significantly less present. Looking closer we can see that this population is not that old and that it has not yet reached the strong decline that starts from age 60. We can also verify this age spread with the interactive map:

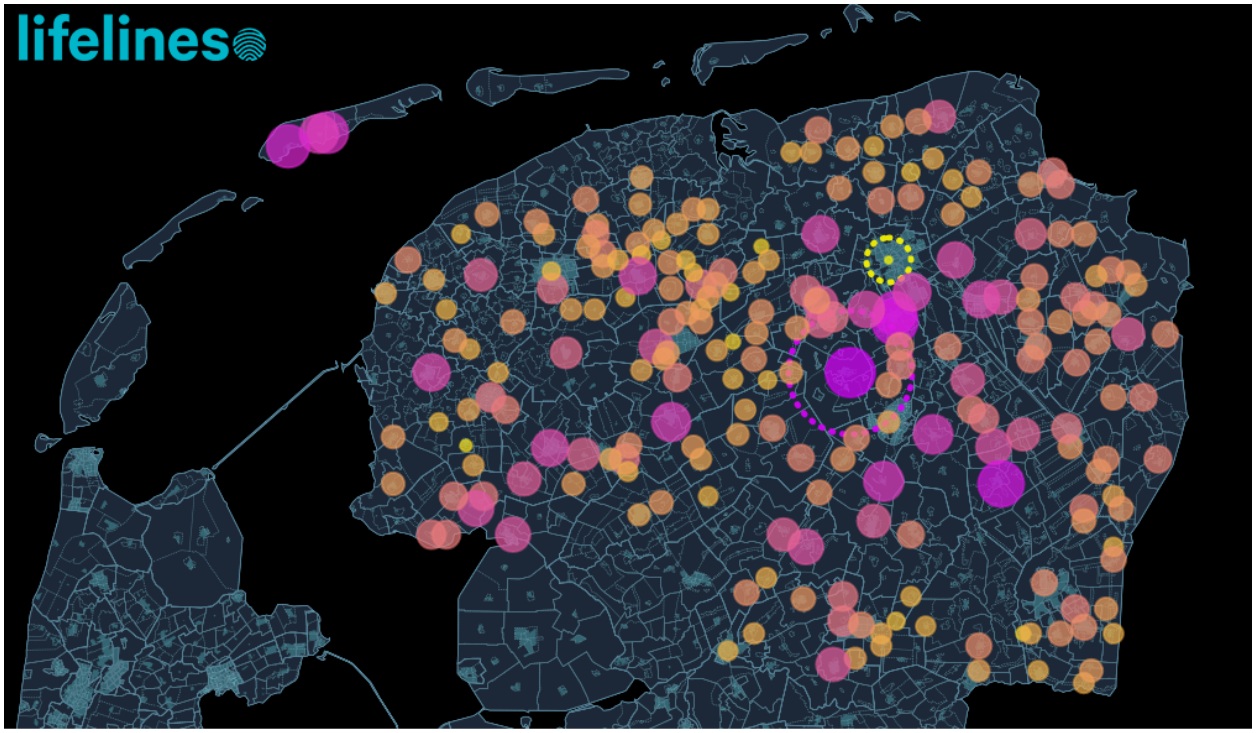


Figure 11: Age distribution on the interactive map

Conclusion

Based on this quick analysis we've seen that:

1. Joint degradation and physical activity have gender differences
2. While the eastern part of Groningen/Drenthe exercises less, this may not be age related
3. Successfully promoting physical activity may require a different approach between men and women

By gaining these insights and surfacing relationships between data you will be able to dig deeper into the possible causes of your challenge. You can continue with similar analyses and/or narrow down on the geographical component (and, for example, create plots like figure 8 but differentiated on regions). You can also use this input to start qualitative research and interview others to verify your findings. This is up to you and your project team to work on and we're curious to learn what you might learn!