Using secondary data sources in public health research: Lessons learned based on two practical examples

Numerous separate CSV files, each with tens of thousands of rows of coded data. As a researcher, how do you deal with such complex data from secondary data sources? This document provides insight into practical experiences regarding the access and use of complex data from secondary data sources for public health research. First, a number of generic data protocol elements are discussed that are important when using secondary data sources. Next, for illustrative purposes, two practical examples are presented detailing lessons learned during specific research projects.

Summary

1. Use of secondary data in public health research:
   - Secondary data, sourced from non-academic partners and various pre-existing resources, offers efficiency for real-world intervention evaluations in public health research.
   - While advantageous for saving time and resources, some challenges include restricted data access, lack of influence on outcome variables, and potential changes in data collection methods.
2. Three main steps to take when working with secondary data:
   - Setting up a Data Sharing Agreement (DSA): formal agreements on data sharing purposes, items to share, data format, data usage, privacy, and responsibilities to adhere to data protection principles.
   - Agreeing upon and experimenting with a data format: Discussing and negotiating the desired data format with the data supplier, requesting data samples, and visualizing ideal data formats for analyses.
   - Pre-registration of scientific research hypotheses and analyses: Pre-registration enhances transparency and trustworthiness by publicly sharing analysis plans prior to the study through recognized trial registers or pre-registration services.
3. Lessons Learned from two practical examples:
   - Real-world supermarket experiment:
     - Highlights the time-consuming process of establishing a DSA with a supermarket chain.
     - Emphasizes the importance of creating an application programming interface (API) for categorizing products as 'healthy' or 'unhealthy.'
     - Acknowledges the advantage of data being aggregated by the supermarket but also notes limitations in data granularity and potential time investments.
   - Real-world experiment within cardiac rehabilitation:
     - Demonstrates the importance of applying FAIR principles for data sharing.
     - Also emphasizes the time-consuming process of establishing a legally correct DSA.
     - Advises careful consideration of data format and collaboration with data suppliers to ensure compatibility with statistical analysis programs.

Secondary data in public health research

Secondary data refers to data which is used for different purposes than the original intent of the data collection and is managed by non-academic partners. It can be derived from various pre-existing sources, such as electronic health records, mobile phone apps, or health surveillance data.[1] Using secondary data can be of particular interest for evaluating intervention effects in real-world settings, using real-world outcome measures. For example, the analysis of changes in disease risk markers derived from electronic health records following the implementation of a new treatment. The fact that secondary data are collected for other purposes has the major advantage of saving time and resources for the researcher using the data, but also introduces challenges. These challenges can be huge, such as restricted data access, lack of influence on the outcome variables of interest and their measurement methods, and potential changes in data collection methods or used coding of the data over time.[2] Goal of this document is to provide researchers with an overview of lessons learned in dealing with complex secondary data in practice, to facilitate other researchers with getting access to and the use of secondary data in public health research. But before this, we will start by discussing the three main steps to take when the aim is to work with secondary data for scientific research purposes.

Setting up a data sharing agreement

A first step, when aiming to use secondary data sources in public health research, is establishing a formal data sharing agreement (DSA) with the secondary data supplier. This also facilitates adhering to data protection principles. In the DSA, agreements can be made on the purpose of data sharing, the exact data items and data format to be shared, how the data will be used by the data receiver, how data privacy is secured, and what the roles and responsibilities are of each party involved (e.g., removal of the data after a set time frame). Regarding the use of the data by the data receiver, agreements can be made on the use of data for scientific publications, including no interference with the analysis or interpretation of results by the data supplier, and the right to publish no matter the outcomes.

Agreeing upon and experimenting with the data format

As a secondary data receiver, you typically have no influence on the data collection methods or units of measurement. However, you may request to receive the data in a certain format or at a certain level of aggregation. It can be helpful the visualise a data format which would be ideal for conducting your analyses, and discuss possibilities with the data supplier how to best match this desired format in the data exchange. It may be that the data supplier can extract their data on a different aggregation level, which could closely match your desired format, without the need for (much) extra work by the data supplier. It can also be helpful to request a data sample in advance, in order to explore data distributions and finalise a statistical analyses plan. Such a sample will also provide detailed insight into the data format to be expected and for example the coding used (if no codebook is available), and it can be used for testing of data transformations and the drafting of a detailed plan on how most efficiently transform these data into the desired data format.

Pre-registration of the scientific research hypotheses and analyses

Pre-registration of an analysis plan enables research transparency and good conduct by publicly sharing the analyses plans prior to conducting the study, enhancing the trustworthiness and replicability of study findings. Pre-registration can be approached in various ways which are not necessarily mutually exclusive. Popular options are (1) registration in a WHO-recognised trial register (which ideally also allows uploading of analyses plans, such as the ISRCTN registry), (2) via publication of a full study protocol in a peer-reviewed journal, or (3) via registration in a pre-registration service (e.g., OSF registry). The pre-registration service is especially useful for registration of nonclinical studies

[1] Näher AF, et al. Secondary data for global health digitalization. Lancet Digit Health 2023;5:e93–101.
[2] Boslaugh S. Using Secondary Data. Public Health Research Methods. 2015; SAGE Publications, Inc. [ISBN 9781483398839]

and for secondary analyses of trials including human participants which are already registered in a trial register.

Getting data access

Data access can be organised by establishing methods for data exchange, again taking into account data protection. For example, data can be shared via a secured platform or a virtual research environment (such as anDREa)[3] at which data can be uploaded by the data supplier and from which it can be downloaded and (when desired) can stored at a secured institute network drive by the data user. Furthermore, the timeline for data exchange should be established, and potential related costs should be agreed upon.

Practical example 1: Real-world supermarket experiment

Project description

The first example relates to the Supreme Nudge supermarket experiment testing the effect of a combined nudging and pricing intervention, promoting healthy foods and beverages on individual-level outcomes (e.g., dietary intake and food purchases) and supermarket-level outcomes (e.g., supermarket sales data). This study was conducted according to a cluster-randomised controlled trial design. Further details about the design of this study can be found elsewhere[4]. This project used a combination of primary and secondary data sources, in which the individual-level outcomes were the primary data collected by the research team and the supermarket-level outcomes were the secondary data collected by the supermarket.

The research question addressed in the current practical example is to what extent nudging and pricing strategies changed supermarket level sales trends during the intervention period. Outcome measures were defined as the total percentage of healthy food and beverage purchases and the total percentage of healthy purchases within various food groups, analysed via controlled interrupted time series analyses. For each outcome, we tested for changes in the average sales trends over time in intervention supermarkets compared to the control supermarkets, adjusting for the pre-intervention period sales trends and clustering of data within supermarkets.

Lessons learned

Data sharing agreement and data access

A DSA was established at the start of this project. It described Amsterdam UMC as receiving institute by the external data supplier – the supermarket chain. Our experience is that establishing a legally correct DSA is a time consuming process, which should be started as soon as possible and multiple months should be allocated for completion of the whole process. The main delaying factor is the required evaluation of the DSA by legal and privacy experts of each party involved, generally leading to the exchange of multiple versions before a final agreement is reached.

Data format

We aimed to promote healthy food and beverage purchases and subsequently analyse the percentage of healthy sales as study outcome. It was thus necessary to categorise all supermarket products into healthy versus unhealthy products, while such a categorisation is not regular practice for supermarkets. Therefore, prior to intervention implementation, we organised creation of an application programming interface (API) by the Netherlands Nutrition Centre via which the supermarket chain could extract weekly updated data on which products were recommended within

[3] LinkedIn.com/company/andrea-consortium

[4] Stuber JM, et al. Reducing cardiometabolic risk in adults with a low socioeconomic position: protocol of the Supreme Nudge parallel cluster-randomised controlled supermarket trial. Nutr J. 2020;19,46.

the Dutch dietary guidelines ('healthy') or were not recommended ('unhealthy'). Creating this API was a highly time consuming process, for which multiple months should be allocated and financial budget is required. Yet, ultimately, it saved a lot of work since supermarket assortments are dynamic (hundreds of product numbers change on a weekly basis), requiring frequent updates. Product categorization by us as researchers would thus have been very time consuming.

Raw supermarket sales data constitutes of single transactions per product number for each supermarket location, covering already thousands of daily transactions. As researchers, we were interested in sales data of our 12 participating supermarkets over a one year intervention period plus a six month pre-intervention period. Requesting the raw sales data was therefore considered not feasible. We visualized our ideal data format, and discussed with the supermarket chain which level of data aggregation would be feasible for them to deliver. Based on the API, indicating which products are healthy versus unhealthy, and on existing product categorisation groups by the supermarket, supermarkets employees were able to extract data files detailing per supermarket, per week, the total number of products sold for a certain food or beverage category (e.g., fresh fish), divided by healthy versus unhealthy products. The supermarket chain shared a data sample of this data format, based on a single supermarket, in order to check if this format indeed met our needs.

Pre-registration

Based on the received data format sample, we finalized a pre-registration of these analyses in an online pre-registration service[5]. The pre-registration service was deemed most appropriate, since an original trial protocol was already published on the individual-level outcomes,. The trial protocol indicated that supermarket-level data will be analysed, but without any specific details on how this would be approached.

Data transformations

The final dataset resulted in a single Excel file consisting of approximately 104,000 rows, which is very manageable. The supermarket chain also shared data on which products were precisely categorized in their food and beverage categories. They distinguished 63 categories, which we reassigned to 10 overarching food groups which were relevant for our outcome analyses (fruits, vegetables, legumes, and nuts; grain products; milk and yogurt products; cheese; meat products, meat substitutes and eggs; fish, oils, fats, and herbs and spices; beverages; non-alcoholic beverages; products from all remaining food products (e.g., pre-packaged meals, and baking products); and sweet and savoury snacks). We calculated the sum of healthy and of unhealthy sales of each of these 10 overarching food groups per week per store via the Excel PivotTable function. Based on these data, we were able to calculate the percentage healthy sales per food group and in total. Next, we used an R script to recode all week numbers in the sales data to create an equal time variable for each supermarket indicating the pre-intervention week numbers (week 1 until 26) and intervention weeks (week 27 until 78). Last, a group allocation variable (control or intervention supermarket) and an interruption moment variable (0 = week 1 until 26, 1 = week 27 until 78) were added to finalise the dataset for analyses. Some visual examples of these data transformations can be found in Appendix 1.

The fact that the supermarket chain already aggregated the data on a higher level has been a tremendously time saving approach. The disadvantage was that we were bound to the 63 food groups the supermarket used, due to which we for example were unable to distinguish nuts from vegetables (as those were partly categorised in the same food group). Another disadvantage was the fact that we requested time investment from the supermarket employees, which was again a time consuming process of multiple months before we received the final dataset.

[5] Stuber JM, et al. Pre-registration: Controlled interrupted time series analysis of store-level sales data as part of the Supreme Nudge randomised controlled supermarket trial: OSF Preregistration; 2022 [Available from: doi.org/10.17605/OSF.IO/6KTVJ].

Practical example 2: real world experiment within cardiac rehabilitation

Project description

The second example comes from the BENEFIT project, which relates to an experiment taken place within cardiac rehabilitation care. The goal of this project was to promote initiating and maintaining a healthy lifestyle among CVD patients. For this, the BENEFIT intervention was developed as an addition to standard, face-to-face, cardiac rehabilitation (CR) care. Core of the BENEFIT intervention was access to an advanced eHealth platform with a Personal Health Environment (PHE) consisting of functionality for daily goal monitoring, access to (evidence-based) lifestyle interventions, personal coaching and a reward program aimed at stimulating a wide range of health behaviours and therapy adherence. The intervention also promoted self-management of a healthy lifestyle by encouraging patients to measure health indicators (such as blood pressure) themselves at home. This study utilized a cluster non-randomized controlled trial, to examine the added value of the BENEFIT intervention compared to a control condition wherein patients only received regular cardiac rehabilitation care. Primary outcomes were changes in lifestyle behaviours (i.e. in the domain of (1) exercising (2) smoking (3) alcohol use (4) diet (5) stress (6) sleep). Secondary outcomes were physical outcomes such as BMI and waist circumference and motivational outcomes such as motivation for lifestyle change, self-confidence and (subjective) goal achievement. Outcomes were measured directly after CR (after approximately 3-4 months) and after approximately one year after the start of CR.

Lessons learned

Data sharing agreement

In the BENEFIT project, there was a special work package centring around FAIR data. The people involved in this work package thus thought about the application of FAIR (Findable, Accessible, Interoperable and Reusable) principles, patient privacy, and secure data transfer already from the start of the project. There were also scheduled biweekly meetings with IT personnel of the data supplier to discuss this important topic. We agreed upon a technical infrastructure, based on the anDREa digital research environment, to share data without compromising patient privacy and local data sharing policies. Based on these agreements, an official data sharing agreement (DSA) was drawn up between Leiden University as the research institute and CardioVitaal as the data sharing health centre. In the background, other parties also had to be involved in this matter. For example, next to the research project leaders, also the university's research desk for policy support was involved as well as a privacy officer. At the side of the health centres, the DSA was made possible with the chief executive officer (CEO) and the chief operating officer (COO) of CardioVitaal together with the ICT department at Vital10. Thus, establishing a legally correct DSA takes much deliberation, counsel, discussion and evaluation by different parties and thus is a time consuming process.

In addition, when you want to use data from patients, it's necessary to ask permission for your research idea from a medical ethics committee. The committee reviews the research protocols and provides feedback to ensure human safety and research quality. For example, in these protocols, it should be made clear that the patients from which you want to use the data that has been collected, need to have explicitly consented to letting their information be used for research purposes by secondary partners.

Pre-registration

At the start of the project, we pre-registered our project within the 'Nationaal Trial Register' (NTR, now 'Landelijk Trial Register' (LTR)). This was a preregistration of the whole project, centring around the aims of the project, general hypotheses and outcomes. In addition, for every separate research paper, we are in the process of registering our specific hypotheses, our selection of variables necessary to test these hypotheses and according specific data-analyses techniques in OSF 'Open Science Framework' Registries. OSF Registry provides a transparent and easily accessible repository that provides an easy-to-use format for study preregistrations.

Data format and data transformations

At first, the data format was discussed only very generally between the BENEFIT project manager and IT personnel from the data supplier. The researchers received pilot data after a few months and then found out that it was very hard to transpose and restructure the data in such a way that they could work with it in a number of different statistical analysis programs. For example, we received data files for each separate project[6], and then, for each project, separate files. These files consisted, for example, of participant IDs, Variables IDs, appointment data, questionnaire data etc., resulting in many different data files, with each data file sometimes consisting of more than a million cases (i.e., rows), that had to be combined. We thus made a request for a different data structure and a limiting of the number of different data files to be received. This was luckily possible, but did took the data supplying company almost a year to develop. We thus advice to always ask for a preliminary data dump, to test the possibilities of the data and check the steps that need to be taken to be able to work with the data (i.e., restructuring, aggregating, combining and transforming the data).

As always, but especially with data from external agencies, it is important to carefully check the data. Especially when you as a researcher are only interested in data from a selection of participants (for example, only participants that agreed to provide their data for research purposes, or only from participants receiving a certain intervention, or from participants connected to the company during a certain time frame), it is so important to check whether you receive the data from the correct participants. In our project, we managed to negotiate to also keep track of the influx of patients ourselves. At the health centre, we could see which new patients were scheduled for CR, and we could check whether they were appointed to the right projects[7]. These numbers could later be checked against the (pseudonymised) data received from the data supplier, categorised by project. This way, we could examine whether the numbers matched, and if not, figure out the reason. It turned out that, during the project, a number of small mistakes were made by IT personnel of the data supplier in the encoding of the data dumps, which could be fixed after some detective work. Without the information of the influx of new patients, we would never have found these mistakes. Thus as a researcher, it is so important to always carefully check the data and, if possible, be on top of data collection. Close collaboration with the external data supplier is necessary.

Other 'lessons learned'

One of the best tips for researchers who want to receive and work with complex secondary data is to convince the data supplier to appoint a project manager with whom the researcher can discusses the state of affairs, preferably on a weekly basis. For example, have updates been made concerning the protocol (e.g. protocol to inform new personnel, to contact participants, etc.), and concerning the data (e.g., changes in data format, inclusion of new variables, new response categories, different coding schemes, etc.)? Or changes regarding the control / intervention condition, or in personnel that will affect implementation of the intervention or IT services/data collection? For example, during the BENEFIT project, one CR centre started a new project once, offering the patients new and adjusted (research) questionnaires without our knowledge. Also, coding of responses were sometimes changed by IT personnel without our knowledge (which was especially unfortunate as the same coding could then mean two different things). Finally, personnel changes are also inevitable in long-term projects, thus there was a continuing need to monitor the training of new staff providing the intervention. For the project manager at the side of the researchers, it was vital to be aware of all these changes to (1)

[6] i.e., projects could be seen as patient journeys. Thus with each new project, something changed in the patient journey such as a different welcome message, different information, notifications and questionnaires. Sometimes projects were developed in co-creation with the researchers (i.e. separate projects for the control group and intervention group), but also sometimes projects were changed or developed by the health supplier without the project leaders' or project manager's knowledge.

[7] If not, we could contact IT or the involved healthcare professional to correct the mistakes

note it down in the research log and (2) negotiate solutions to mitigate possible unwanted side effects of decisions, and (3) decide on strategies which are necessary to help with the implementation of complex interventions in a real world setting. Thus, appointing one employee as project manager within the external organisation too (paid by either the project or the external agency) who is aware of everything that is going on and who will acts as the main contact person for the project manager at the researchers' side, is vital for a good working partnership and thereby successful project. Note however that this will mean an extra investment and will thus not always be possible, or only possible for a limited amount of time (e.g. start phase).

Authors
Josine M. Stuber, Amsterdam UMC location Vrije Universiteit Amsterdam, Epidemiology and Data Science
Linda D. Breeman, Leiden University, Health, Medical, and Neuropsychology Unit
Joreintje D. Mackenbach, Amsterdam UMC location Vrije Universiteit Amsterdam, Epidemiology and Data Science

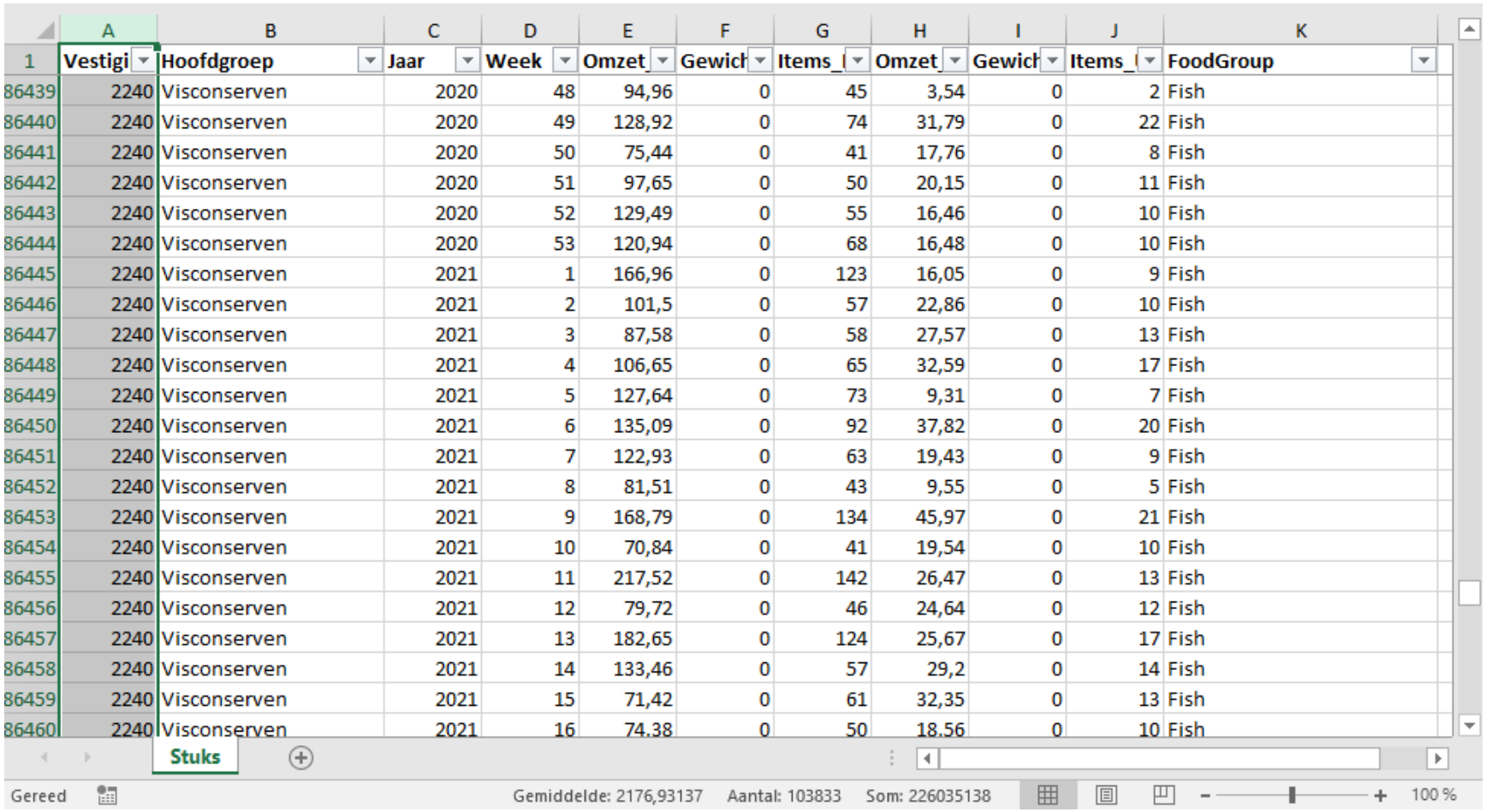
Universiteit Leiden

Amsterdam UMC
Universitair Medische Centra

Appendix 1. Visualisation of data transformations

1. Screenshot of all sales data

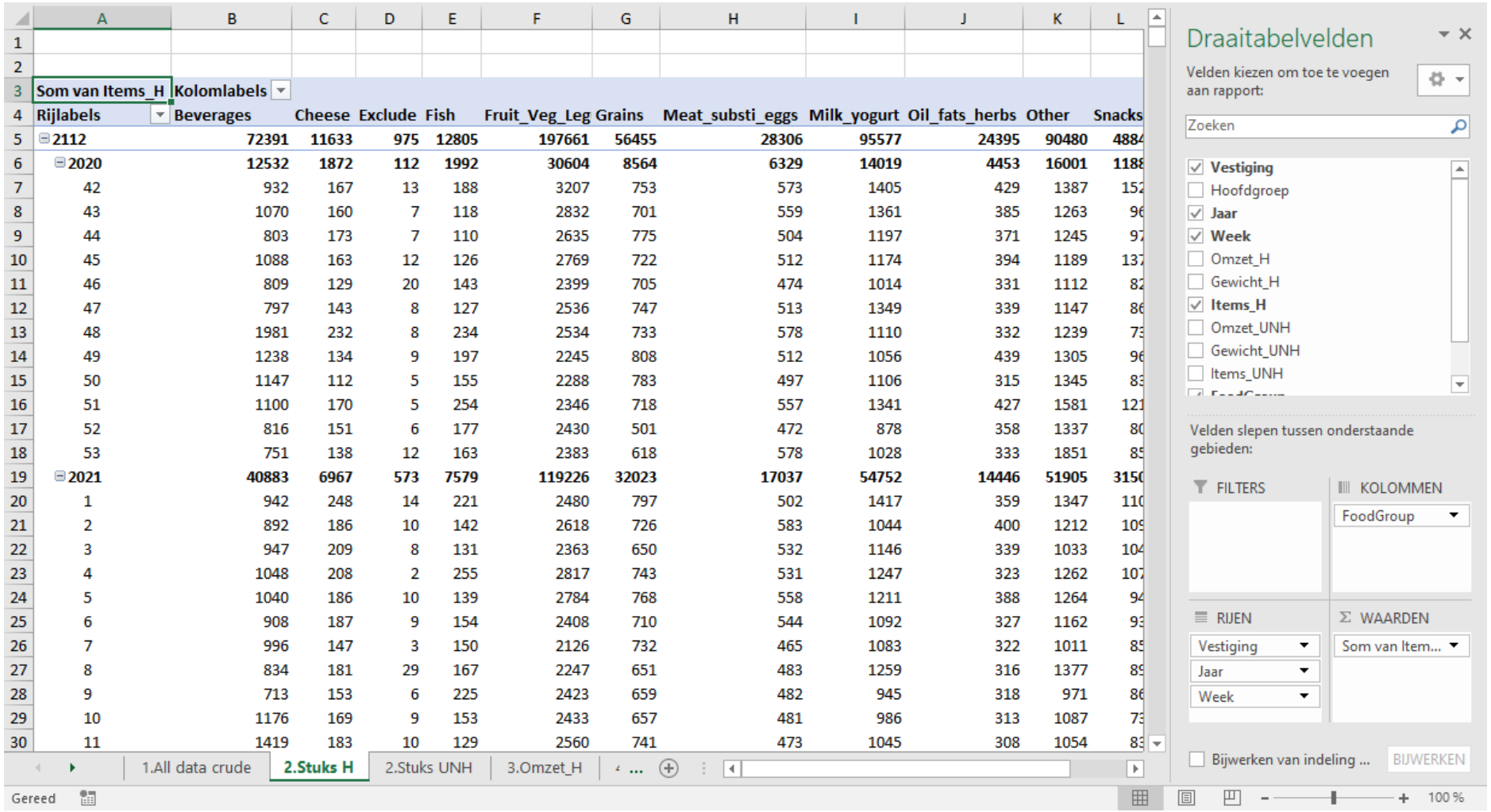
Screenshot of data in items sold, consisting of approximately 104,000 rows, in which 63 food groups (e.g. 'Visconserven', column B) are categorized into 10 overarching food groups (e.g. 'Fish', column K):



	A	B	C	D	E	F	G	H	I	J	K
1	Vestigi	Hoofdgroep	Jaar	Week	Omzet_	Gewich	Items_	Omzet_	Gewich	Items_	FoodGroup
86439	2240	Visconserven	2020	48	94,96	0	45	3,54	0	2	Fish
86440	2240	Visconserven	2020	49	128,92	0	74	31,79	0	22	Fish
86441	2240	Visconserven	2020	50	75,44	0	41	17,76	0	8	Fish
86442	2240	Visconserven	2020	51	97,65	0	50	20,15	0	11	Fish
86443	2240	Visconserven	2020	52	129,49	0	55	16,46	0	10	Fish
86444	2240	Visconserven	2020	53	120,94	0	68	16,48	0	10	Fish
86445	2240	Visconserven	2021	1	166,96	0	123	16,05	0	9	Fish
86446	2240	Visconserven	2021	2	101,5	0	57	22,86	0	10	Fish
86447	2240	Visconserven	2021	3	87,58	0	58	27,57	0	13	Fish
86448	2240	Visconserven	2021	4	106,65	0	65	32,59	0	17	Fish
86449	2240	Visconserven	2021	5	127,64	0	73	9,31	0	7	Fish
86450	2240	Visconserven	2021	6	135,09	0	92	37,82	0	20	Fish
86451	2240	Visconserven	2021	7	122,93	0	63	19,43	0	9	Fish
86452	2240	Visconserven	2021	8	81,51	0	43	9,55	0	5	Fish
86453	2240	Visconserven	2021	9	168,79	0	134	45,97	0	21	Fish
86454	2240	Visconserven	2021	10	70,84	0	41	19,54	0	10	Fish
86455	2240	Visconserven	2021	11	217,52	0	142	26,47	0	13	Fish
86456	2240	Visconserven	2021	12	79,72	0	46	24,64	0	12	Fish
86457	2240	Visconserven	2021	13	182,65	0	124	25,67	0	17	Fish
86458	2240	Visconserven	2021	14	133,46	0	57	29,2	0	14	Fish
86459	2240	Visconserven	2021	15	71,42	0	61	32,35	0	13	Fish
86460	2240	Visconserven	2021	16	74,38	0	50	18,56	0	10	Fish

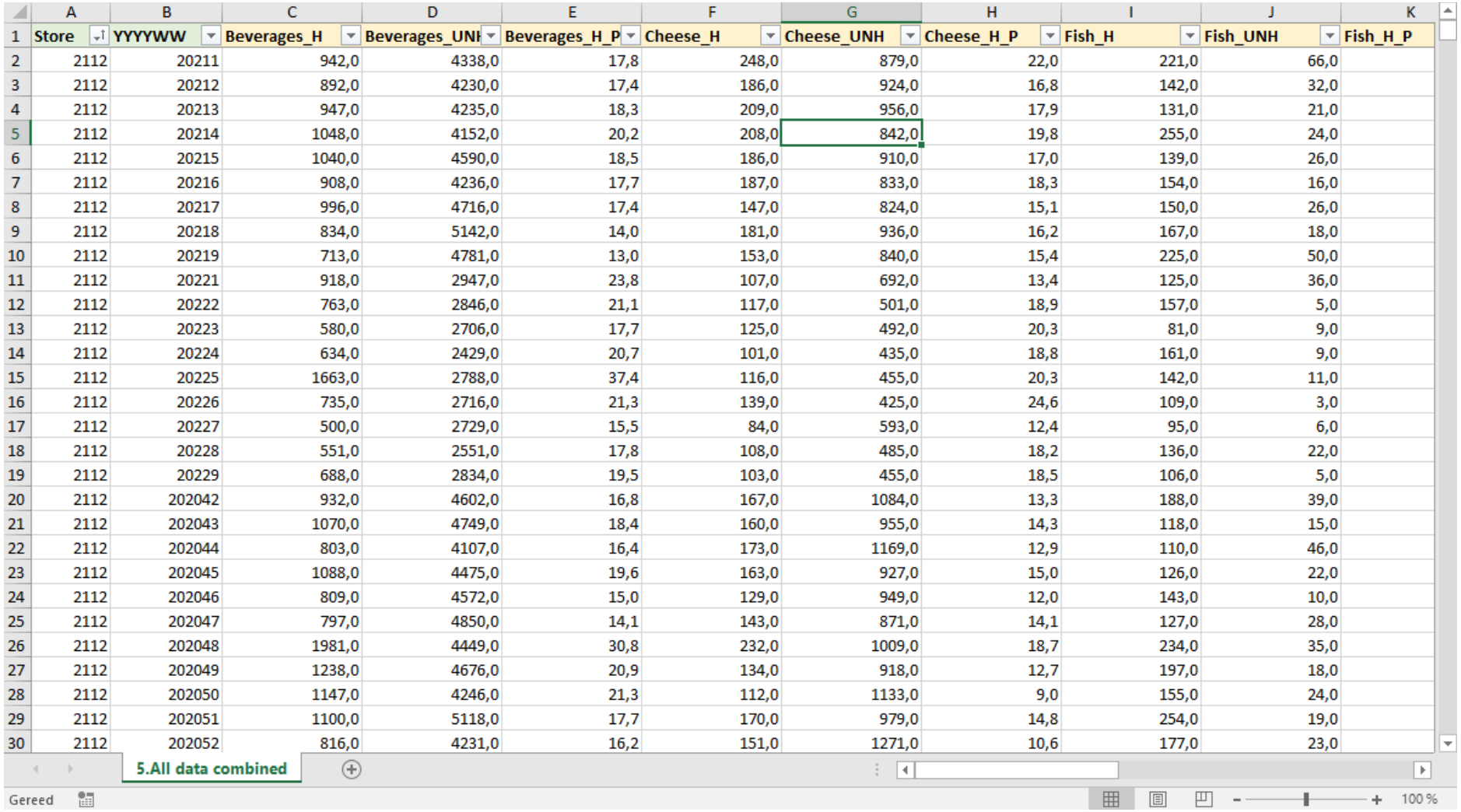
2. Use of Excel PivotTable function to calculate sum of healthy and of unhealthy sales for each food group per week per store

Screenshot of one out of four PivotTables used:



Som van Items_H	Kolomlabels											
Rijlabels	Beverages	Cheese	Exclude	Fish	Fruit_Veg_Leg	Grains	Meat_substi_eggs	Milk_yogurt	Oil_fats_herbs	Other	Snacks
2112	72391	11633	975	12805	197661	56455	28306	95577	24395	90480	4884
2020	12532	1872	112	1992	30604	8564	6329	14019	4453	16001	1188
42	932	167	13	188	3207	753	573	1405	429	1387	152
43	1070	160	7	118	2832	701	559	1361	385	1263	96
44	803	173	7	110	2635	775	504	1197	371	1245	97
45	1088	163	12	126	2769	722	512	1174	394	1189	137
46	809	129	20	143	2399	705	474	1014	331	1112	82
47	797	143	8	127	2536	747	513	1349	339	1147	86
48	1981	232	8	234	2534	733	578	1110	332	1239	73
49	1238	134	9	197	2245	808	512	1056	439	1305	96
50	1147	112	5	155	2288	783	497	1106	315	1345	83
51	1100	170	5	254	2346	718	557	1341	427	1581	121
52	816	151	6	177	2430	501	472	878	358	1337	80
53	751	138	12	163	2383	618	578	1028	333	1851	85
2021	40883	6967	573	7579	119226	32023	17037	54752	14446	51905	3150
1	942	248	14	221	2480	797	502	1417	359	1347	110
2	892	186	10	142	2618	726	583	1044	400	1212	109
3	947	209	8	131	2363	650	532	1146	339	1033	104
4	1048	208	2	255	2817	743	531	1247	323	1262	107
5	1040	186	10	139	2784	768	558	1211	388	1264	94
6	908	187	9	154	2408	710	544	1092	327	1162	93
7	996	147	3	150	2126	732	465	1083	322	1011	85
8	834	181	29	167	2247	651	483	1259	316	1377	89
9	713	153	6	225	2423	659	482	945	318	971	86
10	1176	169	9	153	2433	657	481	986	313	1087	73
11	1419	183	10	129	2560	741	473	1045	308	1054	83

Screenshot of combined sales dataset, based on the four PivotTables:



	A	B	C	D	E	F	G	H	I	J	K
1	Store	YYYYWW	Beverages_H	Beverages_UNH	Beverages_H_P	Cheese_H	Cheese_UNH	Cheese_H_P	Fish_H	Fish_UNH	Fish_H_P
2	2112	20211	942,0	4338,0	17,8	248,0	879,0	22,0	221,0	66,0	
3	2112	20212	892,0	4230,0	17,4	186,0	924,0	16,8	142,0	32,0	
4	2112	20213	947,0	4235,0	18,3	209,0	956,0	17,9	131,0	21,0	
5	2112	20214	1048,0	4152,0	20,2	208,0	842,0	19,8	255,0	24,0	
6	2112	20215	1040,0	4590,0	18,5	186,0	910,0	17,0	139,0	26,0	
7	2112	20216	908,0	4236,0	17,7	187,0	833,0	18,3	154,0	16,0	
8	2112	20217	996,0	4716,0	17,4	147,0	824,0	15,1	150,0	26,0	
9	2112	20218	834,0	5142,0	14,0	181,0	936,0	16,2	167,0	18,0	
10	2112	20219	713,0	4781,0	13,0	153,0	840,0	15,4	225,0	50,0	
11	2112	20221	918,0	2947,0	23,8	107,0	692,0	13,4	125,0	36,0	
12	2112	20222	763,0	2846,0	21,1	117,0	501,0	18,9	157,0	5,0	
13	2112	20223	580,0	2706,0	17,7	125,0	492,0	20,3	81,0	9,0	
14	2112	20224	634,0	2429,0	20,7	101,0	435,0	18,8	161,0	9,0	
15	2112	20225	1663,0	2788,0	37,4	116,0	455,0	20,3	142,0	11,0	
16	2112	20226	735,0	2716,0	21,3	139,0	425,0	24,6	109,0	3,0	
17	2112	20227	500,0	2729,0	15,5	84,0	593,0	12,4	95,0	6,0	
18	2112	20228	551,0	2551,0	17,8	108,0	485,0	18,2	136,0	22,0	
19	2112	20229	688,0	2834,0	19,5	103,0	455,0	18,5	106,0	5,0	
20	2112	202042	932,0	4602,0	16,8	167,0	1084,0	13,3	188,0	39,0	
21	2112	202043	1070,0	4749,0	18,4	160,0	955,0	14,3	118,0	15,0	
22	2112	202044	803,0	4107,0	16,4	173,0	1169,0	12,9	110,0	46,0	
23	2112	202045	1088,0	4475,0	19,6	163,0	927,0	15,0	126,0	22,0	
24	2112	202046	809,0	4572,0	15,0	129,0	949,0	12,0	143,0	10,0	
25	2112	202047	797,0	4850,0	14,1	143,0	871,0	14,1	127,0	28,0	
26	2112	202048	1981,0	4449,0	30,8	232,0	1009,0	18,7	234,0	35,0	
27	2112	202049	1238,0	4676,0	20,9	134,0	918,0	12,7	197,0	18,0	
28	2112	202050	1147,0	4246,0	21,3	112,0	1133,0	9,0	155,0	24,0	
29	2112	202051	1100,0	5118,0	17,7	170,0	979,0	14,8	254,0	19,0	
30	2112	202052	816,0	4231,0	16,2	151,0	1271,0	10,6	177,0	23,0	

3. Recoding of all week numbers in the sales data to create an equal time variable per study supermarket location

Screenshot of recoding schema for week numbers in the raw data (YYWWWW) to equal time points in weeks per study supermarket (T1 t/m T78):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**YYYYWW**	**2240**	**2199**	**2146**	**2186**	R code	**2166**	**2265**	**2179**	**2279**	R code	**2182**	**2158**	**2115**	**2125**	R code
2	202042	1	1	1	1	"202042"="1",	99999	99999	99999	99999	"202042"="99999",	99999	99999	99999	99999	"202042"="99999",
3	202043	2	2	2	2	"202043"="2",	99999	99999	99999	99999	"202043"="99999",	99999	99999	99999	99999	"202043"="99999",
4	202044	3	3	3	3	"202044"="3",	99999	99999	99999	99999	"202044"="99999",	99999	99999	99999	99999	"202044"="99999",
5	202045	4	4	4	4	"202045"="4",	1	1	1	1	"202045"="1",	99999	99999	99999	99999	"202045"="99999",
6	202046	5	5	5	5	"202046"="5",	2	2	2	2	"202046"="2",	99999	99999	99999	99999	"202046"="99999",
7	202047	6	6	6	6	"202047"="6",	3	3	3	3	"202047"="3",	99999	99999	99999	99999	"202047"="99999",
8	202048	7	7	7	7	"202048"="7",	4	4	4	4	"202048"="4",	99999	99999	99999	99999	"202048"="99999",
9	202049	8	8	8	8	"202049"="8",	5	5	5	5	"202049"="5",	99999	99999	99999	99999	"202049"="99999",
10	202050	9	9	9	9	"202050"="9",	6	6	6	6	"202050"="6",	99999	99999	99999	99999	"202050"="99999",
11	202051	10	10	10	10	"202051"="10",	7	7	7	7	"202051"="7",	99999	99999	99999	99999	"202051"="99999",
12	202052	11	11	11	11	"202052"="11",	8	8	8	8	"202052"="8",	99999	99999	99999	99999	"202052"="99999",
13	202053	12	12	12	12	"202053"="12",	9	9	9	9	"202053"="9",	99999	99999	99999	99999	"202053"="99999",
14	20211	13	13	13	13	"20211"="13",	10	10	10	10	"20211"="10",	99999	99999	99999	99999	"20211"="99999",
15	20212	14	14	14	14	"20212"="14",	11	11	11	11	"20212"="11",	99999	99999	99999	99999	"20212"="99999",
16	20213	15	15	15	15	"20213"="15",	12	12	12	12	"20213"="12",	99999	99999	99999	99999	"20213"="99999",
17	20214	16	16	16	16	"20214"="16",	13	13	13	13	"20214"="13",	99999	99999	99999	99999	"20214"="99999",
18	20215	17	17	17	17	"20215"="17",	14	14	14	14	"20215"="14",	99999	99999	99999	99999	"20215"="99999",
19	20216	18	18	18	18	"20216"="18",	15	15	15	15	"20216"="15",	99999	99999	99999	99999	"20216"="99999",
20	20217	19	19	19	19	"20217"="19",	16	16	16	16	"20217"="16",	99999	99999	99999	99999	"20217"="99999",
21	20218	20	20	20	20	"20218"="20",	17	17	17	17	"20218"="17",	99999	99999	99999	99999	"20218"="99999",
22	20219	21	21	21	21	"20219"="21",	18	18	18	18	"20219"="18",	99999	99999	99999	99999	"20219"="99999",
23	202110	22	22	22	22	"202110"="22",	19	19	19	19	"202110"="19",	99999	99999	99999	99999	"202110"="99999",
24	202111	23	23	23	23	"202111"="23",	20	20	20	20	"202111"="20",	99999	99999	99999	99999	"202111"="99999",
25	202112	24	24	24	24	"202112"="24",	21	21	21	21	"202112"="21",	99999	99999	99999	99999	"202112"="99999",
26	202113	25	25	25	25	"202113"="25",	22	22	22	22	"202113"="22",	99999	99999	99999	99999	"202113"="99999",
27	202114	26	26	26	26	"202114"="26",	23	23	23	23	"202114"="23",	99999	99999	99999	99999	"202114"="99999",
28	202115	27	27	27	27	"202115"="27",	24	24	24	24	"202115"="24",	99999	99999	99999	99999	"202115"="99999",
29	202116	28	28	28	28	"202116"="28",	25	25	25	25	"202116"="25",	99999	99999	99999	99999	"202116"="99999",
30	202117	29	29	29	29	"202117"="29",	26	26	26	26	"202117"="26",	99999	99999	99999	99999	"202117"="99999",
31	202118	30	30	30	30	"202118"="30",	27	27	27	27	"202118"="27",	1	1	1	1	"202118"="1",
32	202119	31	31	31	31	"202119"="31",	28	28	28	28	"202119"="28",	2	2	2	2	"202119"="2",
33	202120	32	32	32	32	"202120"="32",	29	29	29	29	"202120"="29",	3	3	3	3	"202120"="3",
34	202121	33	33	33	33	"202121"="33",	30	30	30	30	"202121"="30",	4	4	4	4	"202121"="4",
35	202122	34	34	34	34	"202122"="34",	31	31	31	31	"202122"="31",	5	5	5	5	"202122"="5",
36	202123	35	35	35	35	"202123"="35",	32	32	32	32	"202123"="32",	6	6	6	6	"202123"="6",
37	202124	36	36	36	36	"202124"="36".	33	33	33	33	"202124"="33".	7	7	7	7	"202124"="7".

R-script to recode the week numbers (YYWWWW) to equal time points per week for each study supermarket:

```
#Load data
setwd("file path")
library(readxl)
data <- read_excel("document_name_combined sales data")

#Trial phase 1 supermarket locations: 2240 (Supermarket_1), 2199 (Supermarket_2), 2186 (Supermarket_3), 2146 (Supermarket_4)
data_Phase_1 <- subset(data, Store==2240 | Store==2199 | Store==2186 | Store==2146)
nrow(data_Phase_1)
data_Phase_1$Phase <- "Stores_phase_1"

# Trial phase 2 supermarket locations: 2166 (Supermarket_45), 2265 (Supermarket_6), 2179 (Supermarket_7), 2279 (Supermarket_8)
data_Phase_2 <- subset(data, Store==2166 | Store==2265 | Store==2179 | Store==2279)
nrow(data_Phase_2)
data_Phase_2$Phase <- "Stores_phase_2"

# Trial phase 3 supermarket locations: 2182 (Supermarket_9), 2158 (Supermarket_10), 2115 (Supermarket_11), 2125 (Supermarket_12)
data_Phase_3 <- subset(data, Store==2182 | Store==2158 | Store==2115 | Store==2125)
nrow(data_Phase_3)
data_Phase_3$Phase <- "Stores_phase_3"

data_Phase_1$YYYYWW <- as.factor(data_Phase_1$YYYYWW)
data_Phase_2$YYYYWW <- as.factor(data_Phase_2$YYYYWW)
data_Phase_3$YYYYWW <- as.factor(data_Phase_3$YYYYWW)

library(plyr)
data_Phase_1$YYYYWW <- revalue(data_Phase_1$YYYYWW,
c("202042"="1", "202043"="2", "202044"="3", "202045"="4","202046"="5", "202047"="6", "202048"="7", "202049"="8", "202050"="9", "202051"="10", "202052"="11",
"202053"="12","20211"="13", "20212"="14", "20213"="15", "20214"="16","20215"="17","20216"="18",
"20217"="19","20218"="20","20219"="21","202110"="22","202111"="23","202112"="24","202113"="25","202114"="26","202115"="27","202116"="28","202117"="29","202118"=
"30","202119"="31","202120"="32","202121"="33","202122"="34","202123"="35","202124"="36",
"202125"="37","202126"="38","202127"="39","202128"="40","202129"="41","202130"="42","202131"="43","202132"="44","202133"="45","202134"="46","202135"="47","20213
6"="48","202137"="49","202138"="50","202139"="51","202140"="52","202141"="53","202142"="54","202143"="55","202144"="56","202145"="57","202146"="58","202147"="59
","202148"="60","202149"="61","202150"="62","202151"="63","202152"="64","20221"="65","20222"="66","20223"="67","20224"="68","20225"="69","20226"="70","20227"="71
```

```
","20228"="72","20229"="73","202210"="74","202211"="75","202212"="76","202213"="77","202214"="78","202215"="99999","202216"="99999","202217"="99999","202218"="
99999","202219"="99999","202220"="99999","202221"="99999","202222"="99999","202223"="99999","202224"="99999","202225"="99999",
"202226"="99999","202227"="99999"))

data_Phase_2$YYYYWW <- revalue(data_Phase_2$YYYYWW, c(
"202042"="99999","202043"="99999","202044"="99999","202045"="1","202046"="2","202047"="3","202048"="4","202049"="5","202050"="6","202051"="7","202052"="8","2020
53"="9","20211"="10","20212"="11","20213"="12","20214"="13","20215"="14","20216"="15","20217"="16","20218"="17","20219"="18","202110"="19","202111"="20","202112"
="21","202113"="22","202114"="23","202115"="24","202116"="25","202117"="26","202118"="27","202119"="28","202120"="29","202121"="30","202122"="31","202123"="32","
202124"="33","202125"="34","202126"="35","202127"="36","202128"="37","202129"="38","202130"="39","202131"="40","202132"="41","202133"="42","202134"="43","20213
5"="44","202136"="45","202137"="46","202138"="47","202139"="48","202140"="49","202141"="50","202142"="51","202143"="52",”202144"="53","202145"="54","202146"="55
","202147"="56","202148"="57","202149"="58","202150"="59","202151"="60","202152"="61","20221"="62","20222"="63","20223"="64","20224"="65","20225"="66","20226"="6
7","20227"="68","20228"="69","20229"="70","202210"="71","202211"="72","202212"="73","202213"="74","202214"="75","202215"="76","202216"="77","202217"="78","202218
"="99999","202219"="99999","202220"="99999","202221"="99999","202222"="99999","202223"="99999","202224"="99999","202225"="99999","202226"="99999","202227"="9
9999",
"202041"="99999","202040"="99999","202039"="99999","202038"="99999","202037"="99999","202036"="99999","202035"="99999","202034"="99999","202033"="99999","202
032"="99999","202031"="99999","202030"="99999","202029"="99999","202028"="99999"))

data_Phase_3$YYYYWW <- revalue(data_Phase_3$YYYYWW, c(
"202042"="99999","202043"="99999","202044"="99999","202045"="99999","202046"="99999","202047"="99999","202048"="99999","202049"="99999","202050"="99999","202
051"="99999","202052"="99999","202053"="99999","20211"="99999","20212"="99999","20213"="99999","20214"="99999","20215"="99999","20216"="99999","20217"="99999
","20218"="99999","20219"="99999","202110"="99999","202111"="99999","202112"="99999","202113"="99999","202114"="99999","202115"="99999","202116"="99999","2021
17"="99999","202118"="1","202119"="2","202120"="3","202121"="4","202122"="5","202123"="6","202124"="7","202125"="8","202126"="9","202127"="10","202128"="11","202
129"="12","202130"="13","202131"="14","202132"="15","202133"="16","202134"="17","202135"="18","202136"="19","202137"="20","202138"="21","202139"="22","202140"="
23","202141"="24","202142"="25","202143"="26","202144"="27","202145"="28","202146"="29","202147"="30","202148"="31","202149"="32","202150"="33","202151"="34","20
2152"="35","20221"="36","20222"="37","20223"="38","20224"="39","20225"="40","20226"="41","20227"="42","20228"="43","20229"="44","202210"="45","202211"="46","2022
12"="47","202213"="48","202214"="49","202215"="50","202216"="51","202217"="52","202218"="99999","202219"="99999","202220"="99999","202221"="99999","202222"="9
9999","202223"="99999","202224"="99999","202225"="99999","202226"="99999","202227"="99999"))

#Combine datasets
StoreLevelData <- rbind(data_Phase_1, data_Phase_2, data_Phase_3)
```

4. Finalize dataset for analyses

R-script to add a group allocation variable (control or intervention supermarket) and an interruption moment variable:

```
library(dplyr)
StoreLevelData$Time <- as.numeric(StoreLevelData$Time)
data <- rename(StoreLevelData, Time="YYYYWW")

#Add group variable
setwd("file path")
Group_code_per_store <- read_excel("Group code per store.xlsx")
StoreLevelData <- merge(data,Group_code_per_store, by = "Store", all = TRUE)

#Add interruption variable
StoreLevelData$Interruption <- ifelse(StoreLevelData$Time >26, c("Post-intervention"), c("Pre-intervention"))

#Save final dataset for data analyses
library("xlsx")
write.xlsx(StoreLevelData, file = "StoreLevelData.xlsx", sheetName = "StoreLevelData", col.names = TRUE, row.names = TRUE, append = FALSE)
```

Screenshot of final dataset ('StoreLevelData') for analyses:

	A	B	C	D	E	F	G	H	I	J	K
1	Store	Time	Group	Location	Interruption	Beverages_H	Beverages_UNH	Beverages_H_P	Cheese_H	Cheese_UNH	Cheese_H_P
184	2146	12	Control	[redacted]	Pre-intervention	1144	4269	21,1343063	306	1909	13,8148
185	2146	13	Control		Pre-intervention	1178	4404	21,10354712	351	1550	18,4639
186	2146	14	Control		Pre-intervention	1131	4467	20,20364416	249	1337	15,699
187	2146	15	Control		Pre-intervention	1894	3654	34,13842826	315	1593	16,5094
188	2146	16	Control		Pre-intervention	1568	3806	29,1775214	274	1388	16,4861
189	2146	17	Control		Pre-intervention	1487	4472	24,95385132	287	1439	16,6280
190	2146	18	Control		Pre-intervention	1231	3847	24,24182749	284	1347	17,4126
191	2146	19	Control		Pre-intervention	1284	4515	22,14174858	243	1300	15,748
192	2146	20	Control		Pre-intervention	1156	5352	17,76275353	264	1565	14,43
193	2146	21	Control		Pre-intervention	1039	4346	19,29433612	265	1354	16,3681
194	2146	22	Control		Pre-intervention	1567	3926	28,52721646	290	1609	15,2711
195	2146	23	Control		Pre-intervention	2653	3740	41,498514	299	1473	16,8735
196	2146	24	Control		Pre-intervention	1130	4524	19,98585073	273	1279	17,5902
197	2146	25	Control		Pre-intervention	1426	5297	21,210769	247	1536	13,8530
198	2146	26	Control		Pre-intervention	1237	4562	21,33126401	219	1327	14,1655
199	2146	27	Control		Post-intervention	1503	4688	24,27717655	300	1386	17,7935
200	2146	28	Control		Post-intervention	2101	4444	32,10084034	205	1304	13,5851
201	2146	29	Control		Post-intervention	1096	4617	19,18431647	256	1414	15,3293
202	2146	30	Control		Post-intervention	1220	4200	22,50922509	296	1727	14,6317
203	2146	31	Control		Post-intervention	1341	4372	23,47278138	258	1300	16,5596
204	2146	32	Control		Post-intervention	2551	4885	34,30607854	227	1456	13,4878

StoreLevelData