Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2024;16(2):118-28

REVIEW DERLEME

DOI: 10.5336/biostatic.2024-102184

Unlocking the Potential of National Healthcare Data in Türkiye: A Journey Through Challenges and Opportunities: Traditional Review

Türkiye'nin Ulusal Sağlık Verilerinin Potansiyelini Açığa Çıkarma Yolundaki Zorluklar ve Fırsatlar: Geleneksel Derleme

Mehmet KOÇAK[a], Zeynep KÖMBE ELAZAB[b,c]

[a]İstanbul Medipol University International Faculty of Medicine, Department of Biostatistics and Medical Informatics, İstanbul, Türkiye
[b]Boğaziçi University Institute of Social Sciences, Masters Program of Psychological Sciences, İstanbul, Türkiye
[c]Boston University, PhD Program of Psychological and Brain Sciences, Boston, USA

ABSTRACT In this paper, we share our experience between 2020-2023 through our TÜBİTAK BIDEB-2232 International Fellowship for Outstanding Researchers (Award No: 118C306) project titled "Feasibility Assessment and Utility of Combining Streaming National Healthcare Data with Environmental and Food Intake Data to Improve Health Policy and Outcomes". In our feasibility assessment, national data sources covering health and disease outcomes, behavioral data e.g., smoking and environmental factors (e.g., air quality these primary data dimensions are the minimum of what would be needed to initiate any academic pursuits or health policy development efforts. In this regard, we discuss our experiences with data availability depth, quality, and accessibility, along with providing alternative surrogates necessary. If access to such data is possible, we assess challenges regarding the preparatory steps to bring such data from multiple domains into a form that is clean and analysis-ready. We also discuss the importance of the selection of correct modeling framework when dealing with spatially-structured outcome variables such as province level disease prevalence or behavioral data. We conclude that even though there are present sources of health and environmental data, there are limitations at various stages of data availability, accessibility, integration, and utilization, making analyses that would adequately address important epidemiological research questions for developing relevant public health policies in Türkiye challenging.

Keywords: Turkish national healthcare data; data access; spatial modeling

ÖZET Bu makalede, TÜBİTAK BIDEB-2232 Uluslararası Lider Araştırmacılar Programı (Ödül No: 118C306) kapsamında "Sağlık Politikalarının ve Sağlık Çıktılarının İyileştirilmesinde Ulusal Akışkan Sağlık Verilerini (Streaming Healthcare Data) Çevresel ve Gıda Tüketim Verileri ile Birleştirmenin Fizibilitesinin ve Faydasının Değerlendirilmesi" başlıklı projemizdeki 2020-2023 yılları arasında gerçekleşen deneyimlerimizi paylaşıyoruz. Fizibilite değerlendirmesinin bir parçası olarak, tanı sıklığı ve mortalite gibi sağlık çıktısı verilerini, özellikle sigara ve alkol tüketimi gibi zararlı alışkanlıkları kapsayan davranışsal verileri ve hava kalitesi, meteoroloji, su kalitesi ve toprak kalitesini temsil eden çevresel belirteçleri de içerecek şekilde çok boyutlu ulusal veri kaynaklarını tartışıyoruz. Bu temel veri alanları, herhangi bir akademik çalışmanın başlatılmasında veya sağlık politikalarının geliştirilmesinde ihtiyaç duyulacak verilerinin en azını temsil etmektedir. Bu bağlamda, bu tür verilerin mevcudiyetini, derinlik ve genişliğini, bu tür verilere erişim varlığını ya da erişimin kısıtlamalarının seviyesini ve veri talep sürecini tartışıyoruz. Gerekli veriler mevcut değilse veya erişimleri son derece kısıtlanmışsa, bu tür verilere alternatif teşkil edebilecek verileri (surrogate markers) değerlendiriyoruz. Bu veri alanlarına erişim mümkünse, bu verilerin birden çok kaynaktan alınmaları, veri temizliği aşamaları, verilerin birleştirilmiş veriler olarak analize hazır hâle getirilmesi sürecinin adımlarıyla ilgili zorlukları tartışıyoruz. Ayrıca il sağlık prevalansı veya davranışsal veriler gibi mekânsal olarak yapılandırılmış çıktı değişkenleriyle çalışıldığında, doğru modelleme çerçevesinin seçiminin önemini de tartışıyoruz. Sonuç olarak, sağlık ve çevresel veri kaynaklarının mevcut olduğunu gözlemliyoruz, ancak Türkiye'de önemli epidemiyolojik araştırma sorularını uygun bir şekilde ele alacak analizler için veri erişilebilirliği, erişilebilirlik, entegrasyon ve kullanım aşamalarında çeşitli kısıtlamalar bulunmaktadır. Bu da ilgili kamu sağlığı politikalarının geliştirilmesini zorlaştırmaktadır.

Anahtar kelimeler: Türkiye ulusal sağlık verisi; veri erişimi; mekansal modelleme

TO CITE THIS ARTICLE:
Koçak M, Kömbe Elazab Z. Unlocking the potential of national healthcare data in Türkiye: A journey through challenges and opportunities: Traditional review. Turkiye Klinikleri J Foren Sci Leg Med. 2024;16(2): 118-28.

Correspondence: Mehmet KOÇAK
İstanbul Medipol University International Faculty of Medicine, Department of Biostatistics and Medical Informatics, İstanbul, Türkiye
E-mail: mehmetkocak@medipol.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 13 Feb 2024 ***Received in revised form:*** 10 Jun 2024 ***Accepted:*** 10 Jun 2024 ***Available online:*** 25 Jun 2024

Studying national disease epidemiology is critical to assess the level of success, or lack thereof, of the current healthcare system of a given country and the impact of healthcare policy changes on reducing acute and chronic conditions, increasing the success rate of medical interventions and effective access to the healthcare system. Such a task is generally conducted through cross-sectional family surveys, which are limited by the respondent-provided data, which even though not confirmed medically, are still invaluable to describe the prevalence of a given condition, as the magnitude of the healthcare issue at hand can be adequately assessed. Such studies suffer from the fact that they all are limited to the questions listed in the surveys, which cannot be combined with other important segments of public health data such as access to health care systems, drug utilization, etc.

As digital platforms become commonplace globally and in health care, policy makers and academicians now tap into a new, and more organized and detailed data source called "claims" data. The national healthcare data collected and housed by Turkish Ministry of Health (MH)(e-nabız, https://enabiz.gov.tr/) is a unique example of unified data in that it houses the healthcare data for the entire nation in one system including patient demographics, diagnostics, treatments and follow-ups, prescription drugs, information on healthcare providers, and other relevant information.

The Scientific and Technological Research Council of Türkiye (TÜBİTAK, https://www.tubitak.gov.tr/en) has multiple grant support and scholarship programs including International Fellowship for Outstanding Researchers Programme (2232) directed by Directorate of Science Fellowships and Grant Programmes (BİDEB, https://www.tubitak.gov.tr/en/content-the-scope-of-the-fellowship). In 2019, our project application titled "Feasibility Assessment and Utility of Combining Streaming National Healthcare Data with Environmental and Food Intake Data to Improve Health Policy and Outcomes" was accepted and funded as a 3-year research study. This project was inspired by prior uses of national healthcare data to evaluate disease prevalence, environmental factors, health behaviors, socioeconomic factors, healthcare access, and public health interventions associated with disease prevalence. There are various national healthcare datasets that were utilized for research, such as Korea's Health Insurance Review and Assessment and the French national administrative healthcare database.[1,2] Such databases have been used to analyze resource use for diseases, and the utilization of different medical services for a condition.[1,2] Alongside, there have been efforts to integrate health and environmental information to develop community health policies that address the issues relevant to different local communities. An example is the Centers for Disease Control and Prevention National Environmental Public Health Tracking Network, which has been found to be positively impacting public health actions through improved programs, interventions, and responses to health problems by providing data, expertise, and resources.[3] Even though there has not been a comprehensive initiative to understand the availability, accessibility, and quality of such data in Türkiye, previous researchers have attempted to use different aspects of such data. For example, Ceyhan et al. aimed to analyze retrospective data from the Social Security Institution Medical Messenger (MEDULA) records between 2010 and 2014, focusing on knee arthroplasty surgeries in Türkiye.[4] There have also been geographically small-scale studies on the relationship between health variables and environmental factors.[5,6] However, systematic national data or province level factors are crucial in adequately assessing research questions and developing effective policies.

In this project, we investigated the availability, depth, and access restrictions of national healthcare and related environmental data, alternatives if data is unavailable or highly restricted, and challenges in preparing data from multiple domains for analysis. In this regard, we wanted to assess the utility of the available data in conducting analyses that would address important research questions in epidemiological phenomena and public health issues of Türkiye.

The primary aim of our research protocol were:

(1) Estimate the overall and region-based prevalence of prevalence by seasonal diseases such as influenza, acute and chronic diseases such as cardiovascular diseases, overweight and obesity, diabetes, early child death, child lead poisoning, as well as various hematologic and tumor oncology diagnoses by demographics of interest.

(2) Correlate the disease prevalence with the environmental data, which will be composed of water, soil, and air quality data from the municipalities and other governmental organizations, along with the food intake data will be obtained with the collaborations with the chambers of commerce of local governments.

The above two aims were constructed with the hope of providing a much-needed mechanism to estimate the disease prevalence of a rich array of diagnoses and to identify potential associations of health outcomes, healthcare, and environmental factors. Such findings would allow for more efficient and timely healthcare policies to be developed. There are many examples of studies in the literature that have utilized cross-sectional survey data to investigate national disease epidemiology. An example is Adams and Marano's study on acute conditions, injuries, activity restrictions, chronic conditions, health status, and medical service use, including physician visits and short hospital stays in the non-institutional civilian population of the U.S.[7] Such cross-sectional survey data is crucial for assessing the prevalence of conditions. For instance, Blank et al. reported an increase of based on age in influenza prevalence rates of 37.3% in children under 5 to 46.3% in children aged 5-17, possibly due to school attendance.[8] Such surveys are essential for evaluating the impact of health policies, such as flu vaccination. Similarly, Danaei et al. conducted a meta-analysis of studies totaling up to a sample of 5.4 million individuals from 199 countries, showing a decrease in global systolic blood pressure by about 0.8 mmHg in men and 1.0 mmHg in women per decade.[9]

As an example of utilization of such claims data, Psaty et al. used the US Medicare and Medicaid claims data to investigate cardiovascular health outcomes.[10] Typically, such data sources provide a relational database framework, which allows for the repeated occurrences of health outcomes to be tracked easily. Additionally, information from multiple, potentially independent, data sources can be merged together, allowing for a vast range of health policy and outcome questions to be addressed with far fewer gaps in the available data. However limitations from such "claims" data include incomplete or inaccurate, potentially compromising the reliability and validity of the findings.[11] Moreover, it may be that there exists may be several, totally independent, databased that do not capture the same information of interest in the same way. So, merging such data is not an immediately easy task but requires external laborious work and thus prone to great errors.

CURRENT PROJECT

Here, we share our experience through our BIDEB-2232 research program from 2020 to 2023 with the wider research community in Türkiye and elsewhere; This paper aims to discuss the availability, accessibility, and quality of national healthcare and relevant environmental data, while addressing the strengths and limitations of the data's utilization within public health and epidemiological research. Our investigations have revealed how various limitations in data accessibility, availability, and quality can pose significant challenges to researchers in this field. In this regard, we also highlight potential issues and pitfalls while offering short and long-term solutions to the limitations that are present within the current system.

ACCUMULATING EXPERIENCE

We present our experience under the two main aims of our study: I-Feasibility, II-Utility.

I. FEASIBILITY

Feasibility of Health-Related Data:

Our immediate experience with Turkish healthcare data is that gaining access to the streaming national healthcare data is not as easy as one might think, and it will remain to be difficult for any future research. We requested the healthcare data from Turkish MH on selected diagnoses aggregated at month, year, and province based on the ICD9 codes. Although we requested data summarizing the monthly counts within a given province for years 2017 through 2019, we were told unofficially that MH has other priorities at the moment. Despite our several inquiries regarding these official data requests, we were not granted access to such data.

This gives us the impression that MH is not ready to share even the province level monthly, let alone yearly, data of disease diagnosis. It is discouraging that granular data at patient level, to follow up a patient within the system from diagnosis to prognosis, is far from available. It appears that there is currently no mechanism in place for processing such data requests for external use. As a result, we submitted a proposal outlining data-sharing strategies to the ministry, which has been evaluated at different levels within the ministry (Figure 1). The proposal offers a three-layer data management system within the national healthcare data:

1) Data Pool that is ready to be queried (Data Pool-1): This data pool will consist of there will be clean, verified data fields that are selected and approved by the upper management of the ministry. It will be dynamic pool that can be expanded updated according to the needs of the Ministry as well as based on the requests from external users for academic and other health-related research. Data Pool-1 will be used by MH and its departments only.

2) Data pool that is ready to be used (Data Pool-2): This data pool will be dynamically fed by Data Pool-1 but will anonymize all protected health identifiers such as patient ID, birth date, etc., while allowing a given patient to be followed within the database chronologically with a randomly generated identification number. Data Pool-2 will be used for all external data requests, and it can also serve as a simulation platform for all artificial intelligence and machine learning models. For example, 5% of the nationwide healthcare data can be extracted randomly for researchers as training and test sets for such models.

3) Data pool with readily available data tables (Data Pool-3): This data pool will also be dynamically fed from Data Pool-2 based on common needs of the ministry and will contain short and practical data tables. For example, the Annual Health Statistics report can automatically pool such tables automatically from Data Pool-3. Data Pool-3 can be made public, and any Turkish citizen can access this pool through their e-government account.

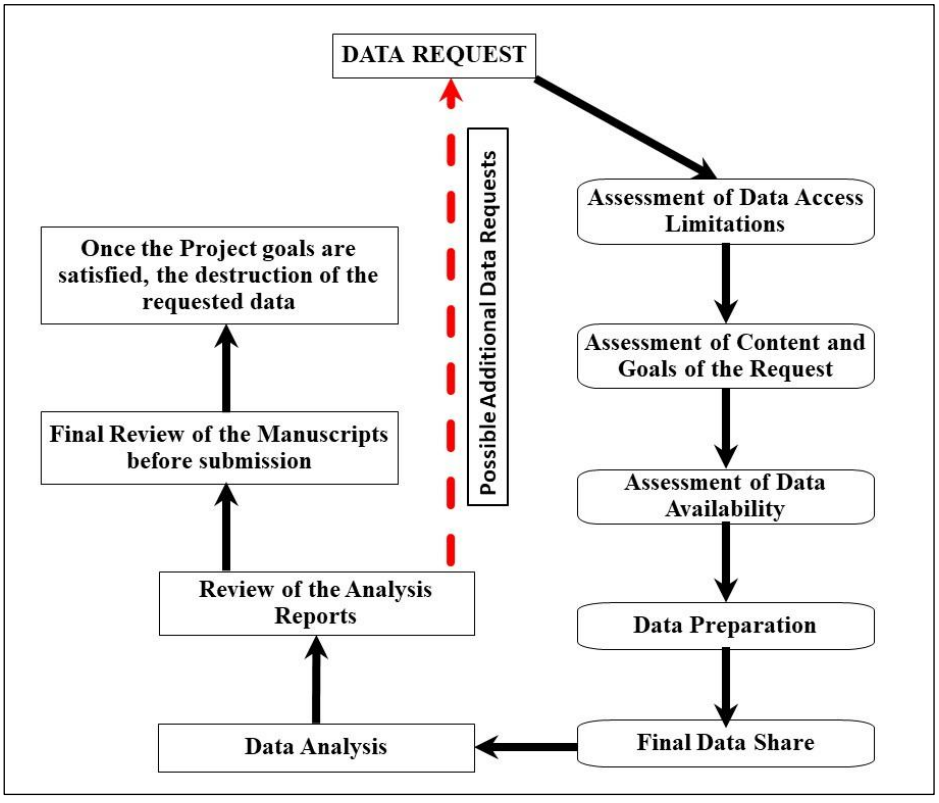

FIGURE 1: Proposed data-sharing structure for Turkish Ministry of Health.

The only data components that we were able to obtain from MH is the province-level data we extracted from the Annual Health Statistics reports (https://sbsgm.saglik.gov.tr/TR-93567/health-statistics-yearbook.html), including markers of access to healthcare. These reports are publicly available both in Turkish and English and contain valuable health-related data.

After experiencing and realizing that we cannot obtain individual-level healthcare data through MH, we redirected our focus towards other sources of aggregated data. For example, there are general reports such as Burden of Disease in Türkiye by Akgun et al. and Turkish Health Research studies conducted by Turkish Statistical Institute [Türkiye İstatistik Kurumu (TÜİK)]; however, these reports are either for Türkiye in general or at the regional level with NUTS1 definition.[12] Although such data does not give us province-level data, we officially requested the Turkish Health Studies (THS) datasets first to see the content of such data and see how it helps our project. Thus, we obtained the THS data for years 2008, 2010, 2012, 2014, 2016, and 2019. Having these datasets does not mean that they can be utilized immediately. Two immediate feasibility challenges are that the variable namesdiffer each year and the datasets are provided in comma-separated values format, without labels or formatted values. Consequently, for each dataset, we must manually extract variable labels and field formats from the data dictionary provided by TÜİK in PDF format. The datasets have varying numbers of fields: THS-2008 has 560 THS-2010 and THS-2012 each have 805 THS-2014 has 476, THS-2016 has 450 and THS-2019 has 295. To increase the utility of these datasets, we must manually work with each of the variables and match the variable across multiple datasets, which is a laborious and honestly an error-prone approach. However, we have no other option and now, our team is currently working on making these datasets ready for analysis.

Feasibility of Environmental Data:

Regarding the environmental data, we aimed to assess the feasibility of obtaining data from soil quality, drinking water quality, air-quality, and meteorology data.

1) Soil Quality: Our discussions with the experts in the Ministry of Forestry and Agriculture showed that there is no systematic province-level soil-quality data. They indicated that the local branches do some testing but mainly for the agricultural purposes rather than health purposes. Our primary need in such soil-quality data was to obtain indications of exposure to harmful substances such as lead, that can only be achieved with systematic data capture in neighborhoods, school grounds, etc. Our general online and literature searches did not lead us to any data sources so we conclude that such endpoints seem to be currently infeasible. Soil tests (http://www.serdalab.com.tr/hizmet-detay-serda-toprak-analizleri-526.aspx) typically include the following markers: soil acidity, arsenic, barium, cadmium, copper, mercury, lead, and zinc. We expect that the soil sample data will be much more sparse both in terms of time and geography while they are expected to be much more stable over time as well Türkiye.

2) Drinking Water Quality: The second feasibility barrier, which still remains as a barrier, is the drinking water quality data. We know that the General Directorate of Public Health has a specific department to monitor and manage the water-quality data from each municipality, and such data is a very good example of streaming data both temporally and spatially. We requested the province-level water quality markers including ammonium, aluminum, pH, nitrate, fluoride, nitrite, copper, selenium, arsenic, cadmium, mercury, antimony, lead, chromium, nickel, boron, bromate as monthly averages. However, we were not granted access to even such aggregate data. We then tried the extract such data from municipality websites by web-scraping techniques, which was quite successful for Ankara but not for other provinces. Wherever available, we downloaded the PDF reports of water-quality tests and tried to extract the data manually. Despite investing substantial time and effort, our hard work yielded only partially representative water-quality data for fewer than 30 provinces, falling far short of the monthly province-level water-quality data needed in our models. So, we conclude that water-quality data for such studies is not feasibly available.

To lessen this data issue, we extracted the province-level drinking water-source data from the Environmental Impact Evaluation (EIE) reports generated by the Turkish Ministry of Environment and Urbanization for years 2000-2018, which included the proportion of drinking water coverage from rivers, dams, lakes, springs, and wells.

3) Air-Quality: This data dimension was surprisingly more widely available to obtain from the publicly available ÇED reports from the Ministry of Environment and Urbanization. The main feasibility challenge was the need to manually extract each air-quality marker, including SO_2, NO_2, O_3, CO, and particulate matter (PM) 10 and 2.5, from PDF reports for 81 provinces, separately for each available year. Even though it was a long-term undertake by our study team, it was definitely worth the effort as these markers are highly critical in many of the statistical models we built. A good example of air quality data is given at https://aqicn.org/city/istanbul/. The air quality data typically includes PM 10 and 2.5 micrometers, ozone, nitrate, sulfur dioxide, and carbon monoxide.

4) Meteorology Markers: We officially requested the meteorology data from the Directorate of Meteorology in the Ministry of Environment, Urbanization and Climate Change and were granted the data on temperature, air pressure, humidity, wind speed, and rainy days at province level. As they shared the data in XLSX format, this data was much easier to prepare for data merging. We conclude that meteorological data is feasibly available for research purposes.

5) Feasibility of Dietary and Food Intake Data: Ministry of Forestry and Agriculture does not have any consumption data on fruits and vegetables, not even on meat consumption. However, Commerce Ministry closely captures the local fruits and vegetables sales. We officially requested province-level fruits and vegetables local-sales data for years 2017, 2018 and 2019, which included 37 different fruits or vegetables to be used as surrogates for the food consumptions. To assess the availability of meat consumption, we communicated iwth the Directorate of Meat and Milk Board and concluded that their data is limited to their 18 stores in 14 provinces, far from having sufficient representation at province-level. We even reached out to İstanbul's fishermen about fish consumption; they indicated that fish catch data is often significantly underreported and may not be reliable.

Despite this gap, even the existing food and vegetable markers we were able to obtain from the Ministry of Commerce have been very instrumental in our model building efforts. Eyles et al. showed significant correlation between the market sales data with actual household nutritional intake.[13] Therefore, we still believe that the local sales data can be reasonable surrogates when individual or household level consumption data are not available.

6) Healthy and Harmful Behavior Data: This dimension of our data need is the most significant. We had an extensive literature search as well as in-depth conversations with MH and TÜİK regarding the lack of province-level harmful behavior data on smoking and alcohol consumption, and there seems to be no easy solution to bridge this data gap. Moreover, as these harmful behavior markers are significant factors for most diseases, the one of the primary reasons for our papers to be rejected is not adequately having these critical control variables in our models. For example, modeling respiratory disease deaths (chronic obstructive pulmonary disease) or lung and throat cancer deaths requires control variables for past or current smoking. Consequently, reviewers rightfully criticize the omission of these key variables, noting that key contributors identified in the literature are not included in the models. To meet this gap albeit in a limited manner, we tried to identify at least the regional smoking and alcohol consumption data and we were only able to obtain such indicators at the NUTS1 regions that are only 12 regions. Building models that cover geographical data from 81 provinces using critical control variables aggregated from just 12 regions is too much aggregation diminishing the accuracy and generalizability of our models. We are currently working on extracting harmful behavior data from the THS as described earlier. While these studies are not designed at the province level but rather at the NUTS2 definition of 26 regions, it is still too much of an aggregation for our models. While better representation than NUTS1. In fact, this is one of the primary reasons why we had to design a province-level national survey to bridge this significant gap. We conclude that without generating national sur-

veys at the provincial representation level, designing studies where such harmful behavior markers are the primary control variables is not feasible.

7) Other Supporting Demographics Data: TÜİK and DataTurkey (https://dataturkey.com.tr/) portals have been very instrumental in obtaining province-level demographics data. Therefore, we can conclude that such socio-demographics data needs can be easily met by what is publicly available at TÜİK data portals and through DataTurkey.

II. UTILITY

Although we were not able to obtain the disease diagnoses data on the diseases that we wanted to study, we were able to obtain access to the deaths due to certain disease categories from TÜİK as described above. We believe that, while not sufficient, these hard endpoints still serve as ideal replacements for our project. We were able to combine these provincial death reasons data in years 2018 and 2019 for the following diseases, including the cancers of bladder, breast, colon, liver, lung, lymphatic system, pancreas, prostate, rectum, stomach, uterus, and respiratory system:

Alzheimer	HIV	Nervous system
Asthma	Heart failure	Pneumonia
Chronic obstructive pulmonary disease	Hepatitis	Prostate hyperplasia
Cerebrovascular	Hypertension	Renal disease
Circulatory system	Injury poisoning	Renal failure
Endocrine system	Ischemic heart	Respiratory system
Epilepsy	Meningitis	Sepsis
Gastroenteritis	Multiple sclerosis	Tuberculosis

We merged these death reasons with the following data dimensions:

Food Consumption: In this dimension, fruit and vegetable sales data covering 37 different fruit or vegetable sold in local markets within each of the 81 provinces was obtained from the Turkish Ministry of Commerce for Years 2018-2020.

Enviornmental Variables: To enrich this dimension, data on *AIR QUALITY* markers, namely, PM 10 and 2.5 (PM_{10} and $PM_{2.5}$), SO_2, CO, NO_2, and O_3 were manually extracted from the periodic EIE reports generated by the Turkish Ministry of Environment and Urbanization for each of the 81 provinces of Türkiye for years 2017-2019. $PM_{2.5}$ was not available for about half of the provinces, so we dropped this marker from our modeling efforts. Province-level radon data was also manually extracted from a radon concentration map of Türkiye by the Turkish Atomic Agency in 2014. Another environment dimension we collected data from was *METEOROLOGY*, where province-level air pressure, humidity, number of rainy days in a year, maximum-average-minimum temperature levels, windspeed, total sunlight, sun radiation, and electromagnetic field. We also aimed to cover the *DRINKING WATER* quality dimension as percentages of drinking water sources from streams, dams, lakes, ponds, springs, and wells. This data was also manually extracted for each of the 81 provinces from the periodic EIE reports generated by the Turkish Ministry of Environment and Urbanization in an annual basis between years 2000-2018.

BEHAVIOR: We were not able to obtain province-level data on harmful behaviors such as alcohol and cigarette consumption. However, such data were available at regional level from 12 regions, and we utilized these control variables although they were far from being representative due to the lack of province-level data. We also obtained the province-level elderly (age>65 years) population proportion as a control variable along with smoking and alcohol consumption.

Data Analysis and Publications

Since death rates from the 81 provinces are expected to be highly spatially correlated, as confirmed by formal statistical testing, ordinary regression models would not be appropriate for modeling these endpoints. This is because the errors of these models are no longer independent, as provinces that are closer to each other would have more similar disease profiles or predictor characteristics, while provinces that are further apart would differ more, as depicted in the following examples. Therefore, we utilized the following spatial models and chose the one that has the best goodness of fit to a given endpoint:

SAR: Spatial Autoregressive Model

SDM: Spatial Durbin Model

SEM: Spatial Error Model

SDEM: Spatial Error Durbin Model

SMA: Spatial Moving Average Model

SDMA: Spatial Durbin Moving Average Model

SARMA: Spatial Autoregressive Moving Average Model

SARMA: Spatial Autoregressive Moving Average Model

SAC: Spatial Autoregressive Confused Model

SDAC: Spatial Autoregressive Confused Model

LINEAR: Linear Model

SLX: Linear model with spatial lag of X (SLX) effects

To fit these models, we used SAS® Version 9.4 (Cary, North Carolina, USA). These models take into account the spatial autocorrelation in the outcome variable as well as in the predictors themselves or both.

In our current publications and conference presentations, we incorporated factors associated with the disease of interest such as smoking, alcohol consumption, and elderly age population, and subsequently assessed whether environmental or dietary markers exhibit significance beyond these established factors. One of the main justifiable criticisms of our manuscripts and the reasons for rejection by the journal reviewers, as discussed in the "feasibility" section, is that we lack properly representative control variables, such as healthy and harmful habits, at the province level. This need forced us to consider finding ways to bridge this data gap by designing a national survey to collect data on additional demographics, chronic disease prevalence, along with healthy and harmful behavior patterns that are representable at the province-level. Another rightful criticism of our manuscripts by the journal reviewers is that the sales data does not necessarily mean the "consumption" data. Although we added some healthy and harmful dietary habits questions to our national survey study, there is still a wide unmet data gap regarding the dietary behavior data. These feasibility issues overall directly impact the level of utility of healthcare data as we have experienced so far as they make generating results and health-policy suggestions that are generalizable and reproducible more difficult if not impossible.

As products of the utility assessments, we have written 9 manuscripts in our projects, of which four has been published and five are under review, and our team also had 6 presentations in conferences (3 national, 3 international). To date, we trained one doctorate student, two masters students, four TÜBİTAK-STAR trainees, and two additional students (one local at İstanbul Medipol University, one international at the University of Tennessee in USA) as collaborators.

DISCUSSION AND FUTURE DIRECTIONS

CHALLENGES AND PITFALLS

As discussed above, one of the main challenges beyond the lack of available data was regarding cooperating with other state apparatus. The idea of collecting, organizing, and analyzing data, with the aim of creating a systematic platform where researchers can apply to access data is still a new concept in Türkiye. While some municipalities might be rigorously collecting various environmental data, this might not be in the priority of other municipalities. Even the data collected by the Turkish Ministry of Environment and Urbanization, which is a central state apparatus, was not uniform across 81 provinces, in which not all air quality markers were assessed across the 81 provinces. This problem, along with MH's claim that they had other priorities than arranging and granting us access to their data, highlights a very salient and fundamental challenge for such projects. The importance given to such data-driven methods to better a nation's policies necessitate organized and systemic approaches to collecting, analyzing, and sharing data. It's crucial for state apparatus to recognize and prioritize data-driven methods, although this recognition may not yet be widespread. Such prioritizing must lead policy-makers to initiate plans regarding new data collecting and sharing methods. Without such a momentum, future projects might come across similar challenges.

DATA CONFIDENTIALITY AND SECURITY

An important issue that arises when using any form of sensitive data is confidentiality and protection of individuals' private health information. It is important to note that any form of heath data that we have received were all de-identified. In the case in which there would be the ability of the researcher to identify a participant, such data were not provided. For example, with regards the data we received on provincial-level data of deaths due to certain diseases we were not given any individual-level information. The data we received was in aggregate form. Furthermore, data for a province was not given if it had fewer than three deaths to prevent any possibility of tracing data to individuals. Such attempts on behalf of state apparatus or other institutions that are responsible for national healthcare data are essential for the rights of individuals.

FUTURE PLANS

We will continue our efforts in expanding our database both in terms of coverage and depth and our efforts on publication and conference presentations. The most significant expansion of our database will be the national survey we conducted to bridge at least some of the data gaps listed in detail above within both the feasibility and utility discussions. Below, we discuss the details of this national survey of 46,000 participants:

1. Demographics including age, gender, education, marital status, employment status, income, family structure, health insurance status, etc.
2. Anthropometrics measures such as height and weight
3. Owning or renting the primary residence
4. Heating method during winter months
5. Car ownership
6. Daily stress and happiness levels
7. Several items capturing habits and behavior covering sleeping, daily breakfast, water consumption, exercise, diet, salt consumption, eating late at night, etc.
8. Smoking and alcohol consumption status
9. Health conditions

- Cardiovascular diseases

- Hypertension
- High cholesterol
- Allergy
- Respiratory diseases
- Gastrointestinal diseases
- Musculoskeletal diseases
- Auditory and vision problems
- Headache
- Cognitive problems
- Diabetes
- Kidney problems
- Osteoporosis

An important point related to our national survey is strengths and limitations of such data in the place of healthcare research. National streaming data, such as the potential data from state systems like "e-nabız", usually has a longitudinal nature, in which they keep track of each patient's health indicators and treatments across years. However, surveys such as the one we have conducted are cross-sectional, which not only limits the descriptive information we can attain, but also the inferential analysis and interpretations we can do. Furthermore, national streaming health data consists of information that patients/individuals may not be fully knowledgeable of (e.g. their specific diagnosis, medication, medication dosage) when answering survey questions. Nevertheless, such surveys are of utmost importance as well since they give us the perspective of patients/individuals. Previous healthcare studies highlight the importance of incorporating the subjectivity of patients and have given researchers and policy-makers insights they may not have reached through other methods.[14]

Incorporating Social and Psychological Perspectives: Even though the main variables of interest in this study were environmental exposures, dietary intake, harmful behaviors, and physiological illnesses, we hopefully want to integrate more social factors to our approaches. Social factors include population characteristics of the province, such a literacy rates, employment levels, education levels, and migration have been shown in previous studies to be predictive of well-being.[15] Other social opportunities related to cultural and educational development, along with healthcare are also important structural variables to account for. Moreover, we aim to study predictors of psychological well-being and distress. Many physiological illnesses have been found to be related to psychological well-being or problems.[16] Furthermore, it is essential that health-care policy makers acknowledge the importance of broadening their understanding of health.[17]

CONCLUSION

This innovative project aimed to combine the streaming healthcare data with streaming fruit and vegetable sales data as well as water, air and soil quality data, providing a greater power to investigate the environmental and nutritional effect on disease prevalence in a spatio-temporal manner. Every country has unique combinations of environmental patterns and health-behavior practices, which necessitates each country to conduct such studies to understand how to tailor health policies to different regions to increase the overall well-being of a society. Such studies are not common in Türkiye, but we hope that our project will accelerate future research and facilitate multi-institutional and international collaborations.

Acknowledgement

This study was partially funded by TÜBİTAK Directorate of Science Fellowships and Grant Programmes (BİDEB)-2232 International Fellowship for Outstanding Researchers. We also thank the Turkish Republic Ministry of Forestry and Agriculture, the Directorate of Meteorology, and the Turkish Statistical Institute for data sharing. The opinions raised in this article solely belong to its authors, and do not represent in any shape or form the position of TÜBİTAK and the afore-mentioned institutions which provided data for our research.

Source of Finance

Partial financial support was received from TÜBİTAK Directorate of Science Fellowships and Grant Programmes (BIDEB)-2232 International Fellowship for Outstanding Researchers (Award No: 118C306).

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Mehmet Koçak;* ***Design:*** *Mehmet Koçak;* ***Control/Supervision:*** *Mehmet Koçak;* ***Data Collection and/or Processing:*** *Mehmet Koçak, Zeynep Kömbe Elazab;* ***Analysis and/or Interpretation:*** *Mehmet Koçak, Zeynep Kömbe Elazab;* ***Literature Review:*** *Mehmet Koçak, Zeynep Kömbe Elazab;* ***Writing the Article:*** *Mehmet Koçak, Zeynep Kömbe Elazab; Critical Review: Mehmet Koçak, Zeynep Kömbe Elazab; References and Fundings: Mehmet Koçak.*

REFERENCES

1. Son C, Lim YC, Lee YS, Lim JH, Kim BK, Ha IH. Analysis of Medical Services for Insomnia in Korea: A Retrospective, Cross-Sectional Study Using the Health Insurance Review and Assessment Claims Data. Healthcare (Basel). 2021;10(1):7. [Crossref] [PubMed] [PMC]
2. Cottin V, Larrieu S, Boussel L, Si-Mohamed S, Bazin F, Marque S, et al. Epidemiology, Mortality and Healthcare Resource Utilization Associated With Systemic Sclerosis-Associated Interstitial Lung Disease in France. Front Med (Lausanne). 2021;8:699532. [Crossref] [PubMed] [PMC]
3. Qualters JR, Strosnider HM, Bell R. Data to action: using environmental public health tracking to inform decision making. J Public Health Manag Pract. 2015;21 Suppl 2(Suppl 2):S12-22. [Crossref] [PubMed] [PMC]
4. Ceyhan E, Gursoy S, Akkaya M, Ugurlu M, Koksal I, Bozkurt M. Toward the Turkish National Registry System: A Prevalence Study of Total Knee Arthroplasty in Turkey. J Arthroplasty. 2016;31(9):1878-84. [Crossref] [PubMed]
5. Küçüksümbül A, Akar AT, Tarcan G. Source, degree and potential health risk of metal(loid)s contamination on the water and soil in the Söke Basin, Western Anatolia, Turkey. Environ Monit Assess. 2021;194(1):6. [Crossref] [PubMed]
6. Mentese S, Mirici NA, Otkun MT, Bakar C, Palaz E, Tasdibi D, et al. Association between respiratory health and indoor air pollution exposure in Canakkale, Turkey. Building and Environment. 2015;93:72-83. [Crossref]
7. Adams PF, Marano MA. Current estimates from the National Health Interview Survey, 1994. Vital Health Stat 10. 1995;(193 Pt 1):1-260. [PubMed]
8. Blank PR, Schwenkglenks M, Szucs TD. Influenza vaccination coverage rates in five European countries during season 2006/07 and trends over six consecutive seasons. BMC Public Health. 2008;8:272. [Crossref] [PubMed] [PMC]
9. Danaei G, Finucane MM, Lin JK, Singh GM, Paciorek CJ, Cowan MJ, et al; Global Burden of Metabolic Risk Factors of Chronic Diseases Collaborating Group (Blood Pressure). National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5•4 million participants. Lancet. 2011;377(9765):568-77. [Crossref] [PubMed]
10. Psaty BM, Delaney JA, Arnold AM, Curtis LH, Fitzpatrick AL, Heckbert SR, et al. Study of Cardiovascular Health Outcomes in the Era of Claims Data: The Cardiovascular Health Study. Circulation. 2016;133(2):156-64. [Crossref] [PubMed] [PMC]
11. Stein JD, Lum F, Lee PP, Rich WL 3rd, Coleman AL. Use of health care claims data to study patients with ophthalmologic conditions. Ophthalmology. 2014;121(5):1134-41. [Crossref] [PubMed] [PMC]
12. Akgun SH, Ozkan S, Rajasekharan KN. Burden of disease in Turkey, 2002-2019. Public Health Open Access. 2022;6(1):000199. [Crossref]
13. Eyles H, Jiang Y, Ni Mhurchu C. Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. J Am Diet Assoc. 2010;110(1):106-10. [Crossref] [PubMed]
14. Sullivan M. The new subjective medicine: taking the patient's point of view on health care and health. Soc Sci Med. 2003;56(7):1595-604. [Crossref] [PubMed]
15. Ahnquist J, Wamala SP, Lindstrom M. Social determinants of health--a question of social or economic capital? Interaction effects of socioeconomic factors on health outcomes. Soc Sci Med. 2012;74(6):930-9. [Crossref] [PubMed]
16. Turner AI, Smyth N, Hall SJ, Torres SJ, Hussein M, Jayasinghe SU, et al. Psychological stress reactivity and future health and disease outcomes: A systematic review of prospective evidence. Psychoneuroendocrinology. 2020;114:104599. [Crossref] [PubMed]
17. Leonardi F. The definition of health: towards new perspectives. Int J Health Serv. 2018;48(4):735-48. [Crossref] [PubMed]