BİYOİSTATİSTİK DERNEĞİ 1997

ÇEVRİM İÇİ KONGRE

22. Ulusal ve 5. Uluslararası

BİYOİSTATİSTİK KONGRESİ

28-30 EKİM 2021

DÜZENLEME KURULU

Prof. Dr. Erdem KARABULUT Hacettepe Üniversitesi - Başkan
Prof. Dr. Seval KUL Gaziantep Üniversitesi
Prof. Dr. Mehtap AKÇİL OK Başkent Üniversitesi
Doç. Dr. Derya GÖKMEN Ankara Üniversitesi
Doç. Dr. İlker ÜNAL Çukurova Üniversitesi
Uzm. Ebru ÖZTÜRK Hacettepe Üniversitesi

BİLİMSEL SEKRETERYA

Prof. Dr. Seval KUL
Gaziantep Üniversitesi
Biyoistatistik Anabilim Dalı
GSM : 0 506 281 3875
E-posta : sevalkul@yahoo.com

ORGANİZASYON SEKRETERYASI

Coraltravel

Bilgi İçin : Cahit Sıtkı Bozkurt
GSM : 0 532 302 0848
E-mail : info@biyoistatistikkongresi.net
Telefon : +90 212 247 46 46
Belgegeçer : +90 212 225 72 64

www.biyoistatistikkongresi.net

XXII. ULUSAL
ve
V. ULUSLARARASI
BİYOİSTATİSTİK KONGRESİ
BİLDİRİ ÖZETLERİ KİTABI

Editör
Prof. Dr. Erdem Karabulut

Yardımcı Editörler
Prof. Dr. Seval Kul
Prof. Dr. Mehtap Akçil Ok
Doç. Dr. Derya Gökmen
Doç. Dr. İlker Ünal
Uzm. Ebru Öztürk

28-30 Ekim 2021

XXII. NATIONAL and V. INTERNATIONAL BIOSTATISTICS CONGRESS BOOK of ABSTRACTS

Editor in Chief

Prof. Erdem Karabulut

Associate Editors

Prof. Seval Kul

Prof. Mehtap Akçil Ok

Assoc. Prof. Derya Gökmen

Assoc. Prof. İlker Ünal

Ebru Öztürk, *MSc*

28-30 October 2021

Kongre Onursal Başkanı	
Prof. Dr. Kadir SÜMBÜLOĞLU	H.Ü. Tıp Fak. Biyoistatistik AD E. Öğretim Üyesi/Biyoistatistik Derneği Onursal Başkanı

Kongre Başkanı	
Prof. Dr. Erdem KARABULUT	Hacettepe Üniversitesi

Düzenleme Kurulu

Prof. Dr. Erdem KARABULUT	Hacettepe Üniversitesi	Başkan
Prof. Dr. Seval KUL	Gaziantep Üniversitesi	Bilimsel Sekreterya
Prof. Dr. Mehtap AKÇİL OK	Başkent Üniversitesi	Üye
Doç. Dr. Derya GÖKMEN	Ankara Üniversitesi	Üye
Doç. Dr. İlker ÜNAL	Çukurova Üniversitesi	Üye
Uzm. Ebru ÖZTÜRK	Hacettepe Üniversitesi	Üye

Honorary President of the Congress

Prof. Kadir SÜMBÜLOĞLU	Hacettepe University, Emeritus Faculty Member of Biostatistics / Honorary President of Biostatistics Association

Congress Chair

Prof. Erdem KARABULUT	Hacettepe University

Organizing Committee

Prof. Erdem KARABULUT	Hacettepe Üniversitesi	Chair
Prof. Seval KUL	Gaziantep Üniversitesi	Scientific secretariat
Prof. Mehtap AKÇİL OK	Başkent Üniversitesi	Members
Assoc. Prof. Derya GÖKMEN	Ankara Üniversitesi	Members
Assoc. Prof. İlker ÜNAL	Çukurova Üniversitesi	Members
Ebru ÖZTÜRK, Msc.	Hacettepe Üniversitesi	Members

XXII. Ulusal ve V. Uluslararası Biyoistatistik Kongresi
XXII. National and V. International Biostatistics Congress
28-30 Ekim 2021 Çevrim İçi Kongre – Online Congress

Bilimsel Danışma Kurulu

Prof. Dr. Ahmet ÖZTÜRK	Erciyes Üniversitesi
Prof. Dr. Alan TENNANT	University of Lucerne, Switzerland
Prof. Dr. Atilla Halil ELHAN	Ankara Üniversitesi
Prof. Dr. Ayşe Canan YAZICI GÜVERCİN	İzmir Tınaztepe Üniversitesi
Prof. Dr. Bahar TAŞDELEN	Mersin Üniversitesi
Prof. Dr. Cemil ÇOLAK	İnönü Üniversitesi
Prof. Dr. Erdem KARABULUT	Hacettepe Üniversitesi
Prof. Dr. Ergun KARAAĞAOĞLU	Hacettepe Üniversitesi
Prof. Dr. Ersin ÖĞÜŞ	Başkent Üniversitesi
Prof. Dr. Ertuğrul ÇOLAK	Eskişehir Osmangazi Üniversitesi
Prof. Dr. Ferhan ELMALI	İzmir Katip Çelebi Üniversitesi
Prof. Dr. Fezan ŞAHİN MUTLU	Eskişehir Osmangazi Üniversitesi
Prof. Dr. Gülşah SEYDAOĞLU	Çukurova Üniversitesi
Prof. Dr. Hafize SEZER	Aydın Üniversitesi
Prof. Dr. Handan ANKARALI	Medeniyet Üniversitesi
Prof. Dr. İlker ETİKAN	Yakındoğu Üniversitesi, K.K.T.C.
Prof. Dr. İmran KURT ÖMÜRLÜ	Adnan Menderes Üniversitesi
Prof. Dr. İsmet DOĞAN	Afyonkarahisar Sağlık Bilimleri Üniversitesi
Prof. Dr. Mehtap AKÇİL OK	Başkent Üniversitesi
Prof. Dr. Melody GHAHRAMANİ	Winnipeg Üniversitesi, Canada
Prof. Dr. Meriç YAVUZ ÇOLAK	Başkent Üniversitesi
Prof. Dr. Pınar ÖZDEMİR	Hacettepe Üniversitesi
Prof. Dr. Recai YÜCEL	Temple University, USA
Prof. Dr. Reha ALPAR	Hacettepe Üniversitesi
Prof. Dr. Saim YOLOĞLU	İnönü Üniversitesi
Prof. Dr. Sanjib BASU	University of Illinois, USA
Prof. Dr. Setenay Kevser DİNÇER ÖNER	Osmangazi Üniversitesi
Prof. Dr. Seval KUL	Gaziantep Üniversitesi
Prof. Dr. Sıddık KESKİN	Van Yüzüncü Yıl Üniversitesi
Prof. Dr. Vildan SÜMBÜLOĞLU	Sanko Üniversitesi
Prof. Dr. Yasemin YAVUZ	Ankara Üniversitesi
Prof. Dr. Yusuf ÇELİK	Biruni Üniversitesi
Prof. Dr. Yavuz SANİSOĞLU	Yıldırım Beyazıt Üniversitesi
Prof. Dr. Z. Nazan ALPARSLAN	Çukurova Üniversitesi
Prof. Dr. Zeki AKKUŞ	Dicle Üniversitesi
Doç. Dr. Alex R. de LEON	University of Calgary, Canada
Doç. Dr. Beyza DOĞANAY ERDOĞAN	Ankara Üniversitesi
Doç. Dr. Can ATEŞ	Aksaray Üniversitesi

Doç. Dr. Cengiz BAL	Eskişehir Osmangazi Üniversitesi
Doç. Dr. Deniz SIĞIRLI	Uludağ Üniversitesi
Doç. Dr. Derya GÖKMEN	Ankara Üniversitesi
Doç. Dr. Emre DEMİR	Hitit Üniversitesi
Doç. Dr. Erdal ÇOŞGUN	Microsoft Genetics, USA
Doç. Dr. Gökmen ZARARSIZ	Erciyes Üniversitesi
Doç. Dr. Harika GÖZÜKARA BAĞ	İnönü Üniversitesi
Doç Dr. İlkay DOĞAN	Gaziantep Üniversitesi
Doç. Dr. İlker ÜNAL	Çukurova Üniversitesi
Doç. Dr. Jale KARAKAYA	Hacettepe Üniversitesi
Doç. Dr. M. Agah TEKİNDAL	Selçuk Üniversitesi
Doç. Dr. S. Kenan KÖSE	Ankara Üniversitesi
Doç. Dr. Selçuk KORKMAZ	Trakya Üniversitesi
Doç. Dr. Şengül CANGÜR	Düzce Üniversitesi
Doç. Dr. Umut ÖZBEK	Icahn School of Medicine, Mount Sinai, USA
Doç. Dr. Ünal ERKORKMAZ	Sakarya Üniversitesi
Doç. Dr. Yaşar SERTDEMİR	Çukurova Üniversitesi
Yrd. Doç. Dr. Duchwan RYU	Northen Illinois University, USA

Scientific Advisory Board

Prof. Ahmet ÖZTÜRK	Erciyes University
Prof. Alan TENNANT	University of Lucerne, Switzerland
Prof. Atilla Halil ELHAN	Ankara University
Prof. Ayşe Canan YAZICI GÜVERCİN	İzmir Tınaztepe University
Prof. Bahar TAŞDELEN	Mersin University
Prof. Cemil ÇOLAK	İnönü University
Prof. Erdem KARABULUT	Hacettepe University
Prof. Ergun KARAAĞAOĞLU	Hacettepe University
Prof. Ersin ÖĞÜŞ	Başkent University
Prof. Ertuğrul ÇOLAK	Eskişehir Osmangazi University
Prof. Fezan ŞAHİN MUTLU	Eskişehir Osmangazi University
Prof. Gülşah SEYDAOĞLU	Çukurova University
Prof. Hafize SEZER	Aydın University
Prof. Handan ANKARALI	Medeniyet University
Prof. İlker ETİKAN	Yakındoğu University, K.K.T.C.
Prof. İmran KURT ÖMÜRLÜ	Adnan Menderes University
Prof. İsmet DOĞAN	Afyonkarahisar Health Sciences University
Prof. Mehtap AKÇİL OK	Başkent University
Prof. Melody GHAHRAMANİ	University of Winnipeg, Canada
Prof. Meriç YAVUZ ÇOLAK	Başkent University
Prof. Pınar ÖZDEMİR	Hacettepe University
Prof. Recai YÜCEL	Temple University, USA
Prof. Reha ALPAR	Hacettepe University
Prof. Saim YOLOĞLU	İnönü University
Prof. Sanjib BASU	University of Illinois, USA
Prof. Setenay Kevser DİNÇER ÖNER	Osmangazi University
Prof. Seval KUL	Gaziantep University
Prof. Sıddık KESKİN	Van Yüzüncü Yıl University
Prof. Vildan SÜMBÜLOĞLU	Sanko University
Prof. Yasemin YAVUZ	Ankara University
Prof. Yusuf ÇELİK	Biruni University
Prof. Yavuz SANİSOĞLU	Yıldırım Beyazıt University
Prof. Z. Nazan ALPARSLAN	Çukurova University
Prof. Zeki AKKUŞ	Dicle University
Assoc. Prof. Alex R. de LEON	University of Calgary, Canada
Assoc. Prof. Beyza DOĞANAY ERDOĞAN	Ankara University
Assoc. Prof. Can ATEŞ	Aksaray University
Assoc. Prof. Cengiz BAL	Eskişehir Osmangazi University

Assoc. Prof. Deniz SIĞIRLI	Uludağ University
Assoc. Prof. Derya GÖKMEN	Ankara University
Assoc. Prof. Emre Demir	Hitit University
Assoc. Prof. Erdal Coşgun	Microsoft Genetics, USA
Assoc. Prof. Gökmen ZARARSIZ	Erciyes University
Assoc. Prof. Harika GÖZÜKARA BAĞ	İnönü University
Assoc. Prof. İlkay DOĞAN	Gaziantep Univesity
Assoc. Prof. İlker ÜNAL	Çukurova University
Assoc. Prof. Jale KARAKAYA	Hacettepe University
Assoc. Prof. M. Agah TEKİNDAL	Selçuk University
Assoc. Prof. S. Kenan KÖSE	Ankara University
Assoc. Prof. Selçuk KORKMAZ	Trakya University
Assoc. Prof. Şengül CANGÜR	Düzce University
Assoc. Prof. Umut ÖZBEK	Icahn School of Medicine, Mount Sinai, USA
Assoc. Prof. Ünal ERKORKMAZ	Sakarya University
Assoc. Prof. Yaşar SERTDEMİR	Çukurova University
Assist. Prof. Duchwan RYU	Northen Illinois University, USA

ÖNSÖZ

Değerli Meslektaşlarımız,

Biyoistatistik Derneği olarak 28-30 Ekim 2021 tarihlerinde çevrim içi düzenleyeceğimiz XXII. Ulusal ve V. Uluslararası Biyoistatistik kongresine sizleri davet etmekten büyük mutluluk duyuyoruz.

Tüm dünyayı etkileyen Covid-19 Pandemisi nedeniyle geçen yıl yüz yüze yapılacak olan kongremizi iptal etmiştik. Bu yıl yapılan yaygın aşılama ile pandeminin dünyada ve ülkemizde kontrol altına alınabileceğini düşünmüştük, ancak pandeminin olumsuz etkilerinin sürmesi nedeniyle kongremizi sanal ortamda gerçekleştirmeye karar verdik. Düzenleme kurulu olarak büyük bir heyecanla kongremizi düzenlemeye yoğunlaştık. Geçen dönem çevrim içi yapılan Pazartesi Konuşmaları etkinliğine olan ilgi ve yoğun katılım, kongremizin başarılı geçeceği konusunda bizleri umutlandırdı. Kongremizin ilk gününde kursların sonrasında konferanslar ve sözlü sunumların yer aldığı bir bilimsel program hazırlamayı planlıyoruz.

Siz değerli meslektaşlarımızın desteği ve katkıları ile bilimsel açıdan zengin ve güncel bilgilerin-çalışmaların sonuçlarının paylaşıldığı başarılı bir kongre gerçekleştireceğimizi düşünüyoruz. Birlikteliğimizi pekiştirmek ve gücümüze güç katmak için tüm meslektaşlarımızı kongremizde buluşmaya davet ediyoruz

Saygılarımızla,

Prof. Dr. Erdem Karabulut
Biyoistatistik Derneği Başkanı / Kongre Başkanı

Prof. Dr. Seval Kul
Bilimsel Sekreterya

PREFACE

Dear Colleagues,

We are very pleased to invite you to the XXII. National and V. International Biostatistics Congress, which will be held online by the Biostatistics Association on 28-30 October 2021.

We canceled our face-to-face congress last year due to the Covid-19 Pandemic, which affected the entire world. With widespread vaccination this year, we thought the pandemic could be brought under control in the world and in our country, but due to the pandemic's ongoing negative effects, we decided to hold our congress virtually.
As the organizing committee, we focused our efforts on organizing our congress with great enthusiasm. The interest and intense participation in the Monday Talks event held online last year gave us hope that our congress will be successful. On the first day of our congress, we will organize two courses based on our members' requests. Then, we intend to prepare a scientific program with conferences and oral presentations on the next two days.

We believe that with your help and contributions, we will be able to organize a successful congress in which scientifically rich and up-to-date information and study results will be shared. We invite all of our colleagues to attend our congress in order to strengthen our unity and add to our strength.

Regards,

Prof. Erdem KARABULUT
Biostatistics Association President/Congress Chair

Prof. Seval KUL
Scientific Secretariat

XXII. Ulusal ve V. Uluslararası Biyoistatistik Kongresi
XXII. National and V. International Biostatistics Congress
28-30 Ekim 2021 Çevrim İçi Kongre – Online Congress

BİLİMSEL PROGRAM / SCIENTIFIC PROGRAM

1. Gün / Day 1 (28 Ekim/October): Salon I				**1. Gün / Day 1 (28 Ekim/October): Salon II**			
	13:30	17:00	Kurs 1. AMOS ile Yapısal Eşitlik Modellemeleri Doç. Dr. İlkay DOĞAN Gaziantep Üniversitesi Biyoistatistik AD		13:30	17:00	Kurs 2. SAS programı ile veri analizi Dr. Öğr. Üyesi Tolga KASKATI Lokman Hekim Üniversitesi Biyoistatistik AD
2. Gün / Day 2 (29 Ekim/October): Salon I				**2. Gün / Day 2 (29 Ekim/October): Salon II**			
Saat/Hour			**Sunum Başlığı/Presentation Title**	**Saat/Hour**			**Sunum Başlığı/Presentation Title**
	09:00	**09:30**	**Açılış Konuşması / Opening Speeche, Prof. Dr. Erdem Karabulut, Kongre Başkanı**		09:00	09:30	
Oturum Başkanları / Chairs: Prof. Dr. Seval Kul, Doç. Dr. Derya Gökmen, Doç. Dr. Cengiz BAL				***Oturum Başkanları/Chairs: -Prof. Dr. Atilla Halil ELHAN, Doç. Dr. İlker Ünal, Doç. Dr. Beyza DOĞANAY ERDOĞAN***			
Oturum1 / Session 1 (Sözlü Sunumlar/Oral presantations)	**09:30**	**09:45**	İndirgenmiş (Reduced) Rank Regresyon, Kısmi En Küçük Kareler ve Temel Bileşenler Faktör Analizi Yöntemlerinin Diyet Örüntüsünü Oluşturmada İstatistiksel Performanslarının Karşılaştırılması ***Mehtap AKÇİL OK***	**Oturum1 / Session 1 (Sözlü Sunumlar/Oral presantations)**	**09:30**	**09:45**	Sağkalım Analizinde Derin Öğrenme ***İrem KAR***, Erdal COŞGUN, Atilla Halil ELHAN
	09:45	**10:00**	Sağlık Alanına İlişkin Boylamsal Verilerin Analizinde Kullanılan Üç Yöntemin Karşılaştırılması: ANOVA, GEE, QLS ***Erdoğan ASAR***, Erdem KARABULUT		**09:45**	**10:00**	Kalp Yetmezliği Hastalarında Sağkalım Tahmini İçin Makine Öğrenmesi Modellerinin Performanslarının Karşılaştırılması Ve Sağkalım İle İlişkili En Önemli Faktörlerin Belirlenmesi Rüstem YILMAZ, ***Fatma Hilal YAĞIN***
	10:00	**10:15**	Çok Kategorili Hastalık Durumlarında Tanısal Modele Yeni Bir Belirteç Eklenmesinin Tanı Performansındaki Değişime Etkisinin İncelenmesi ***Hande ŞENOL***, Ergun KARAAĞAOĞLU		**10:00**	**10:15**	Denetimli Makine Öğrenmesi Yöntemleri ile Gebelerde Anemi ve Anemi ile İlişkili Faktörlerin Tespiti ***Rüveyda YILDIRIM***, Mehmet Onur KAYA, Burkay YAKAR, Bilal ALATAŞ
	10:15	**10:30**	Yapısal Eşitlik Modeli ile Uzunlamasına Veri Analizi: Örtük Büyüme Modelleri ve Gerçek Veri Seti Üzerinde Bir Uygulama ***Funda Seher ÖZALP ATEŞ***, Derya GÖKMEN, İpek GÖNÜLLÜ		**10:15**	**10:30**	Geleneksel Makine Öğrenmesi ve Karışık Etkili Makine Öğrenmesi Model Performanslarının Benzetim Çalışması ile Değerlendirilmesi ***Ebru TURGAL***, Beyza DOĞANAY ERDOĞAN
	10:30	**10:45**	Kişilerin Yaşam Doyumları, Stres Düzeyleri,Beslenme Tutumları ve Uyku Sorunları Arasındaki İlişkilerin Analizi **Muharrem Gürleyen GÖK**, Prof.Dr. Erdem KARABULUT		**10:30**	**10:45**	Ordinal Veri Yapısında Bazı Sınıflandırma Yöntemlerinin Simülasyon Aracılığıyla Karşılaştırılması ***Ali Vasfi AĞLARCI,*** Cengiz BAL
	10:45	**11:00**	Soru-Yanıt / QA		**10:45**	**11:00**	Soru-Yanıt / QA

XXII. Ulusal ve V. Uluslararası Biyoistatistik Kongresi
XXII. National and V. International Biostatistics Congress
28-30 Ekim 2021 Çevrim İçi Kongre – Online Congress

Oturum Başkanları / Chairs: Prof. Dr. Ersin ÖĞÜŞ, Doç. Dr. Erdal COŞGUN				***Oturum Başkanları / Chairs: Prof. Dr. Mehtap AKÇİL OK, Doç. Dr. Jale KARAKAYA, Doç. Dr. Can ATEŞ***			
Oturum 2 / Session 2 (Sözlü Sunumlar/Oral presantations)	**11:15**	**11:30**	Predictive Performance of Expectation Maximization, K-Nearest Neighbor and Random Forest Imputation Methods for Propensity Score Matching in Missing Data Problem ***Buğra VAROL,*** İmran KURT ÖMÜRLÜ, Mevlüt TÜRE	**Oturum 2 / Session 2 (Sözlü Sunumlar/Oral presantations)**	**11:15**	**11:30**	Gradyan Artırma ve Hafif Gradyan Artırma Algoritmaları ile Kırım Kongo Verisi Üzerine Uygulaman ***Osman DEMİR***, Yunus Emre KUYUCU
	11:30	**11:45**	Comparison of Different Machine Learning Methods in the Diagnosis of Chronik Renal Failure (CRF) ***Tolga DEMİREL***, Leman TOMAK		**11:30**	**11:45**	RNA-Dizileme Verilerinin Kümelemesinde Özdüzenleyici HaritalarAlgoritmasına Dayalı Yeni Bir Yaklaşım ***Ahu CEPHE***, Necla KOÇHAN, Gözde ERTÜRK ZARARSIZ, Vahap ELDEM, Erdem KARABULUT, Gökmen ZARARSIZ
	11:45	**12:00**	Survival Prediction with Supervised Principal Components Analysis, Penalized Cox Models and Extreme Learning Machine based Penalized Cox Models on Simulated High Dimensional Data ***Fulden CANTAŞ,*** İmran KURT ÖMÜRLÜ, Mevlüt TÜRE		**11:45**	**12:00**	Akciğer, Prostat ve Meme Kanseri Mikrodizi Gen İfade Verilerinde Öznitelik Seçimi ve Sınıflama ***Özlem ARIK***, Erdem KARABULUT
	12:00	**12:15**	Identifying the Factors Affecting the Intensive Care Unit Admission of Covid-19 Patients by Using Multivariate Statistical Methods ***Neslihan GÖKMEN İNAN***, Arzu BAYGÜL EDEN		**12:00**	**12:15**	Çoklu Omik Verilerinin Birleştirilmesinde Kullanılan Yaklaşımların Ve Sınıflandırma Yöntemlerinin Performansının Araştırılması ***Funda İPEKTEN***, Gözde ERTÜRK ZARARSIZ, Halef Okan DOĞAN, Vahap ELDEM, Gökmen ZARARSIZ
	12:15	**12:30**	Semiparametric shared frailty models for right-censored data using an EM algorithm ***Aysun CETINYUREK YAVUZ***, Philippe LAMBERT		**12:15**	**12:30**	Mikrodizi Verilerinde Kullanılan Medyan ve Cyclic Loess Normalleştirme Yöntemlerinin Derin Öğrenme Performanslarına Etkisinin Karşılaştırılması ***Asena Ayça ÖZDEMİR***, Gülhan TEMEL, Saim YOLOĞLU
	12:30	**12:45**	Soru-Yanıt / QA		**12:30**	**12:45**	Soru-Yanıt / QA
Oturum Başkanları/Chairs:Prof. Dr. Vildan SÜMBÜLOĞLU, Doç. Dr. Yaşar SERTDEMİR,Doç. Dr. Pınar GÜNEL KARADENİZ				***Oturum Başkanları/Chairs: Prof. Dr. C. Reha ALPAR, Doç. Dr. Gülhan TEMEL, Doç. Dr. İlkay DOĞAN***			
Oturum 3 / Session 3 (Kısa Sözlü Sunumlar/Short Oral Presentations)	**14:00**	**14:05**	A Nomogram Model to Predict Bone Metastasis in Patients with Breast Cancer ***Senem KOÇ***, Halil İbrahim TANBOGA, Fuzuli TUGRUL	**Oturum 3 / Session 3 (Kısa Sözlü Sunumlar/Short Oral Presentations)**	**14:00**	**14:05**	İstatistik ve Olasılık Dergilerinde Açık Erişim ve Abonelik Erişimli Dergiler: 1999-2018 dönemi için karşılaştırmalı bir inceleme ***İlker ETİKAN***
	14:05	**14:10**	Investigation of the Effect of Hepatosteatosis on Severity Score Obtained by Repeated Measures in COVID-19 Patients ***Şirin ÇETİN***, Özge PASİN, Ahmet Turan KAYA		**14:05**	**14:10**	Avrupa Sağlık Okuryazarlığı Ölçeği Türkçe Versiyonu Kısa Form Önerisi ***Mehmet Ali SUNGUR***, Zerrin GAMSIZKAN, Demet Hanife SUNGUR
	14:10	**14:15**	A Spatial Analysis of Air Quality and Food Consumption Surrogates in Relation to Prostate Cancer Deaths Selman AKTAŞ, ***Mehmet KOÇAK***		**14:10**	**14:15**	Nedensel Aracılık Analizi ***Nihan ÖZEL ERÇEL,*** Hilal ALTUNDAL, Gülhan TEMEL, Mualla YILMAZ, Cemil ÇOLAK
	14:15	**14:20**	Avrupa Birliği Üye Ülkeleri ve Aday Ülke Türkiye'nin Sağlık Göstergeleri Bakımından Çok Değişkenli İstatistiksel Yöntemlerle Analiz Edilmesi ***Anıl TUZCU***, Cengiz BAL		**14:15**	**14:20**	Popülasyon Genetiği Çalışmalarında Mantel Testi Üzerinde bir Uygulama ***Osman Tolga KASKATI***, Mustafa Murat ARAT, Fatma KAYMAKAMTORUNLARI DENİZ, Emre KESKİN

	14:20	**14:25**	Gözlemsel Çalışmalarda Yapılan Meta-Analiz Çalışmalarının Raporlanması MOOSE Kontrol Listesinin Türkçe'ye Uyarlanması ***Arzu Baygül EDEN***, Neslihan Gökmen İNAN		**14:20**	**14:25**	Sentetik Kontrol Yöntemi Nedir Ve Neden Kullanılmalı? ***Özge PASİN***, Handan ANKARALI
	14:25	**14:30**	Örneklemin Evrendeki Oranı ile Güven Aralığı Genişliğinin Uyumsuzluğu Mehmet Güven GÜNVER, ***Aydan MALÇOK***, Uğurcan SAYILI, Mine SEZGİN, Hatice Sena ÖNER, Sevda ÖZEL YILDIZ, Başak GÜRTEKİN, Eray YURTSEVEN		**14:25**	**14:30**	Genel Özellikleri ile Hata Ağacı Analizi Sıddık KESKİN, ***Kübranur TUNÇAY***
	14:30	**14:35**	Kantil Regresyon Ve Doğrusal Regresyon Yönteminin Karşılaştırılması: Bir Simülasyon Çalışması ***Didem ÜNAL***, Gülhan TEMEL				
	14:35	**14:40**	Pediatrik Büyüme Eğrileri İçin Box-Cox Cole Green (BCCG) Dağılımında Parametrelerin Etkili Serbestlik Derecesinin Belirlenmesi ve Düzleştirme Sonucunda Model Performanslarının Karşılaştırılması ***Eda ÇAKMAK***, Ergun KARAAĞAOĞLU				
	14:40	**15:00**	Soru-Yanıt / QA		**14:30**	**15:00**	Soru-Yanıt / QA
Saat/Hour			**Invited talk**	**Saat/Hour**			**no invited talk**
Oturum Başkanları/Chairs: Prof. Dr. Z. Nazan ALPARSLAN							
Invited talk	**15:30**	**16:30**	Bioconductor on the Cloud ***Nitesh Turaga,*** Scientist II, Department of Data Science, at the Dana Farber Cancer Institute of Harvard Medical School&Bioconductor Core Team Member				
3. Gün/Day 3 (30 Ekim/October): Salon I				**3. Gün/Day 3 (30 Ekim/October): Salon II**			
Saat/Hour			**Sunum Başlığı/Presentation Title**	**Saat/Hour**			**Sunum Başlığı/Presentation Title**
*Oturum Başkanları/**Chairs**: **Prof. Dr. İlker ETİKAN, Doç. Dr. Adem DOĞANER***				*Oturum Başkanları/**Chairs**: **Prof. Dr. Yusuf ÇELİK, , Doç. Dr. Özlem KAYMAZ***			
Oturum 4 / Session 4 (Sözlü Sunumlar/Oral presantations)	**09:30**	**09:45**	Sağkalım Tahmininde Kapula Tabanlı Gen Seçimi ***Aslıhan ALHAN***	**Oturum 4 / Session 4 (Sözlü Sunumlar/Oral presantations)**	**09:30**	**09:45**	Box-Cox Dönüşüm Parametresi Kestirimi İçin Yeni Bir Yaklaşım ***Muhammed Ali YILMAZ,*** Osman DAĞ
	09:45	**10:00**	Sağkalım Analizinde Zamana Bağlı Açıklayıcı Değişkenler için Birleşik Model Yöntemi Kullanımı ***Fatma KAYMAKAMTORUNLARI DENİZ***, Osman Tolga KASKATI, Merve BAŞOL GÖKSÜLÜK		**09:45**	**10:00**	Meta Analizinde Yanlılığı Düzeltme Yaklaşımları ***Emre DİRİCAN***
	10:00	**10:15**	Hızlı Biyoistatistik Analizler için Veri Formatı: PARQUET ***Su ÖZGÜR***		**10:00**	**10:15**	Sıralı Yanıt Değişkenleri için Stereotype Lojistik Regresyon ve Sıralı Lojistik Regresyon Modellerinin Karşılaştırılması: Hipermetropi Verileri Üzerine Bir Uygulama ***Hülya ÖZEN***

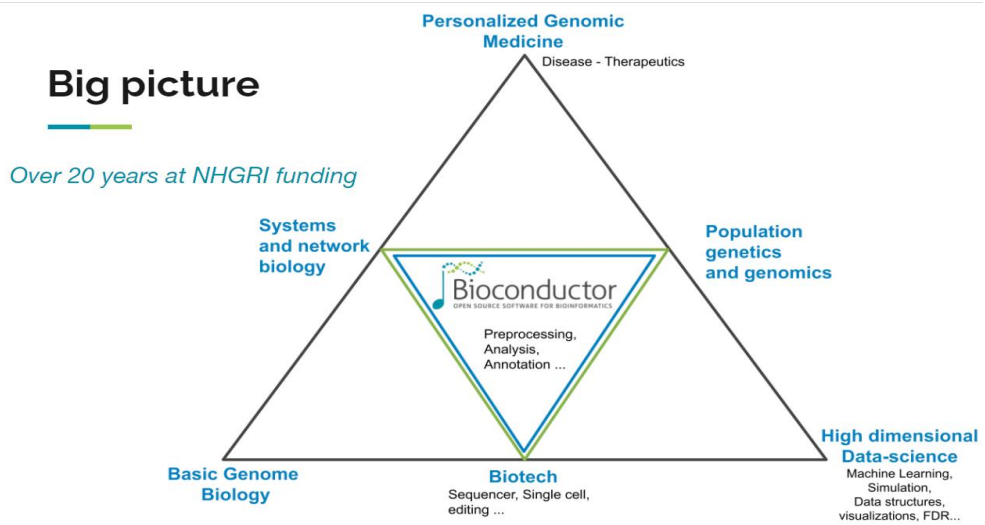
	10:15	**10:30**	Yöntem Karşılaştırma Çalışmalarında En Uygun Regresyon ve Güven Aralığı Yönteminin Belirlenmesine Yönelik Kapsamlı Bir Benzetim Çalışması ***Gözde ERTÜRK ZARARSIZ***, Gökmen ZARARSIZ, Dinçer GÖKSÜLÜK, Cengiz BAL, Halef Okan DOĞAN, Ahmet ÖZTÜRK, Gabi KASTENMÜLLER		**10:15**	**10:30**	Lojistik Regresyon Analizi ile Elde Edilen Beta Katsayısı, Odds Oranı ve Tamsayılı Skor Sisteminin Klinik Tahmin Modeli Geliştirilmesindeki Başarısının Karşılaştırılması ***Gülçin AYDOĞDU***, Emre DEMİR, Yasemin YAVUZ
	10:30	**10:45**	Soru-Yanıt / QA		**10:30**	**10:45**	Soru-Yanıt / QA
*Oturum Başkanları/**Chairs**:**Prof. Dr. Yasemin YAVUZ, Doç. Dr. Emre DEMİR***				*Oturum Başkanları/**Chairs**:**Prof. Dr. Fezan MUTLU, Doç. Dr. Şirin ÇETİN***			
Oturum 5 / Session 5 (Sözlü Sunumlar/Oral presantations)	**11:00**	**11:15**	Seyrek (Sparse) Lojistik Regresyon ve Çok boyutlu Genomik Veri Seti üzerine Bir Uygulama ***Özlem KAYMAZ***, Khaled ALQAHTANİ , Henry WOOD, Arief GUSNANTO	**Oturum 5 / Session 5 (Sözlü Sunumlar/Oral presantations)**	**11:00**	**11:15**	Tip2 Diyabet Hastaları için Yapay Zeka Tabanlı Kişiselleştirilmiş bir Zeki Sistem Önerisi ***Osman Tolga KASKATI,*** Burcu TAŞKAN DEMİRKÖK, Onur YEŞİL, Şeref SAĞIROĞLU
	11:15	**11:30**	Mikrobiyom Verilerinde Kullanılan İstatistiksel Analizler: Obezite İlişkili Mikrobiyom Sekans Verisi Üzerinde bir Uygulama ***Selen YILMAZ IŞIKHAN***, Ceren ÖZKUL		**11:15**	**11:30**	A Composite Ranking of Risk Factors for COVID-19 Time-To-Event Data from a Turkish Cohort ***Ayşe ÜLGEN,*** Hatice DÖRTOK DEMİR, Şirin ÇETİN, Wentian LI.
	11:30	**11:45**	Mamografi Görüntülerinde Ön İşlemede Kullanılan Filtreleme Yöntemleri için Belirlenen Farklı Parametre Değerlerinin Sınıflama Başarısına Etkisi ***Hanife AVCI***, Jale KARAKAYA		**11:30**	**11:45**	Klinik Karar Vermede Sınıflandırma Algoritmalarının Destek olarak Kullanımı: Hipotiroidizm ilişkili Baş Ağrısı Üzerine Bir Uygulama ***Mehmet Ali SUNGUR***, Demet Hanife SUNGUR, Derya ULUDÜZ, Rabia Gökçen GÖZÜBATIK ÇELİK, Esra HATİPOĞLU
	11:45	**12:00**	Çevre Epidemiyolojisi Alanında Kullanılan Zaman Serisi Modellerinin Bibliyometrik ve Altmetrik Skorların Karşılaştırılması ***Mehmet Karadağ***		**11:45**	**12:00**	Gen İfade Veri Setinde Boyut İndirgeme Yöntemlerinin Sınıflama Performansına Etkilerinin Karşılaştırılması ***Fatma Hilal YAĞIN,*** Harika Gözde GÖZÜKARA BAĞ
	12:00	**12:15**	Soru-Yanıt / QA		**12:00**	**12:15**	Soru-Yanıt / QA
	12:30	**13:00**	**Kapanış Konuşmaları / Closing Speeches**				

SÖZLÜ BİLDİRİLER / ORAL PRESENTATIONS

KISA SÖZLÜ BİLDİRİLERİ / SHORT ORAL PRESENTATIONS

Invited Talk

Bioconductor on the cloud
Nitesh Turaga
Scientist II
Bioconductor Core Team



General Update - Bioconductor

[Bioconductor 3.14 is now ready](#) - Oct 27th 2021

R 4.1.1 - *'Kick Things'*

2083 software packages
408 experiment data packages
904 annotation packages
29 workflows
8 books

Books

Books are complex, literate programming works -

- text
- visualizations
- code

Example: Orchestrating Single Cell Analysis - OSCA book

https://bioconductor.org/books/release/

Bioconductor usage

Local

- local computer
- HPC - Linux mostly

Cloud

- Private clouds
- Public clouds

Why the cloud?

Data deluge - size (just lots of data) and type (sparse)

- Single cell - 'sparse' datasets.

Compute power - As much as you need + scalable / elastic

- Scalable compute resources - pay as you go - use what is needed

Pre built tools for users

- Lots of plug and play options

Collaborative

Containers

- Containers are the building blocks to creating end to end cloud solutions.
- Lightweight Virtual machines.
- Bioconductor produces Docker images

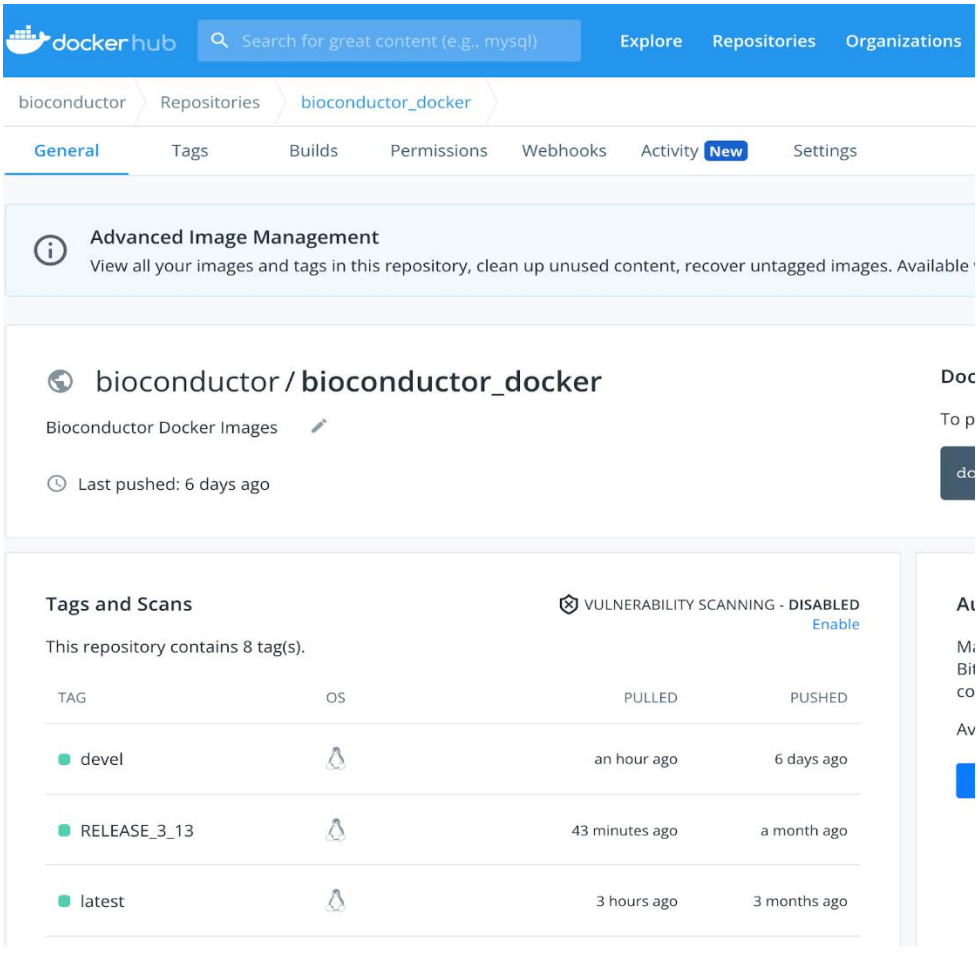
Docker is container engine

Bioconductor Docker images

- **bioconductor/bioconductor_docker**:[devel, RELEASE_X_Y]
  - System dependencies
  - Install > 99% bioconductor packages
  - Ubuntu 20.04
  - RStudio
- **RELEASE_3_14 coming soon!**

http://bioconductor.org/help/docker/

Docker hub

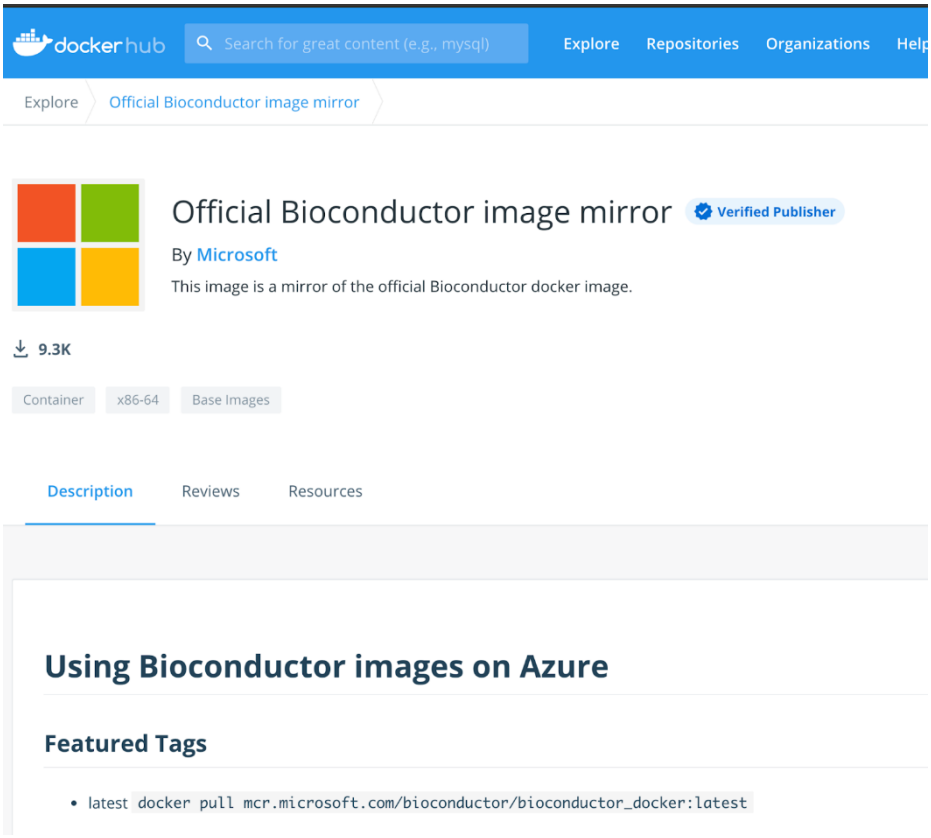


https://hub.docker.com/repository/docker/bioconductor/bioconductor_docker

Usage of Bioconductor Docker images

```
> BiocDockerManager::available()
# A tibble: 6 × 5
  IMAGE                          DESCRIPTION     TAGS      REPOSITORY  DOWNLOADS
  <chr>                          <chr>           <chr>     <chr>           <dbl>
1 bioconductor_docker            Bioconductor D… latest,… bioconduct…    214799
2 bioc-redis                     Bioconductor k… RELEASE… bioconduct…     14132
3 orchestratingsinglecellanalysis This repositor… latest,… bioconduct…       990
4 experimenthub_docker           Experiment Hub  latest    bioconduct…       541
5 annotationhub_docker           Annotation Hub  latest    bioconduct…        47
6 website                        Bioconductor w… latest    bioconduct…        29
```

Microsoft Container Registry (MCR)



https://hub.docker.com/_/microsoft-bioconductor-bioconductor-docker

Docker images are customizable

- Customize docker images to finish all the steps end to end.
- Build on existing docker images

http://bioconductor.org/help/docker/#modify

Bioconductor on Cloud is faster

- Full system dependencies
- Test image for developers
- Binary package installation for Bioconductor packages.

Demo: Benchmark analysis of package installation

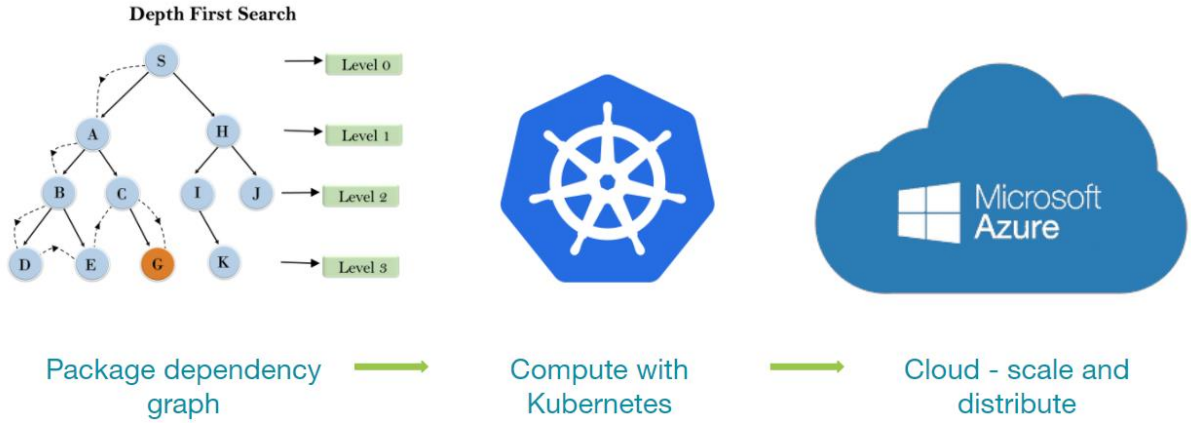
Benchmarking results

```
> res <- microbenchmark::microbenchmark(
   "binary" = {
                  AnVIL::install(c('rhdf5', 'Rhtslib'))
          },
   "traditional" = {
                  BiocManager::install(c('rhdf5', 'Rhtslib'), force = TRUE, ask=FALSE)
          },
          times = 20L
)
```

```
> res
Unit: seconds
        expr       min        lq      mean    median        uq       max neval
      binary  6.681507  6.756271   6.88567  6.916872  6.980436  7.165949    20
traditional 70.776541 71.040255 73.54636 71.410544 71.819983 113.321043    20
```

binary package installation == no compiling packages locally

How are these binaries built



1. Depth First Search package dependency graph.
2. Start bioc-k8s-redis cluster on cloud
3. Build binaries parallely by level in DFS graph
4. Binaries stored on cloud object store.
5. Distribute to users

BiocKubeInstall

Concrete example: OSCA book

```
> pkgs_OSCA_basics_qc <- c('scRNAseq','scuttle', 'robustbase',
                                                    'scater', 'org.Mm.eg.db')
## compare these two!
> AnVIL::install(pkgs_OSCA_basics_qc) # binary install
```

vs

```
> BiocManager::install(pkgs_OSCA_basics_qc) # traditional install
```

Adoption of bioconductor_docker on clouds

Galaxy

RENKU
Collaborative Data Science

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE FOR LIFE SCIENCES TODAY

AnVIL

How can you get started with the cloud?

- Get a cloud account - Azure is a great place to start.
  - Depending on research credits - Google Cloud or AWS are potential solutions as well.
- Look up Bioconductor solutions for cloud
  - Docker images remove configuration issues
  - Binary package installations speed up experimentation / learning

Run Docker images on Azure Container instances

- Easy steps to run these docker images on Azure
- Base image needs to be the bioconductor/bioconductor_docker image.
- Binaries are available for ALL images

Contribution by: Jass Bagga and Dr. Erdal Cosgun

http://bioconductor.org/help/docker/#aci

Use Azure Container Instances to run bioconductor images on-demand on Azure

Azure Container Instances or ACI provide a way to run Docker containers on-demand in a managed, serverless Azure environment. To learn more, check out the documentation here.

Run bioconductor images using ACI

Prerequisites:

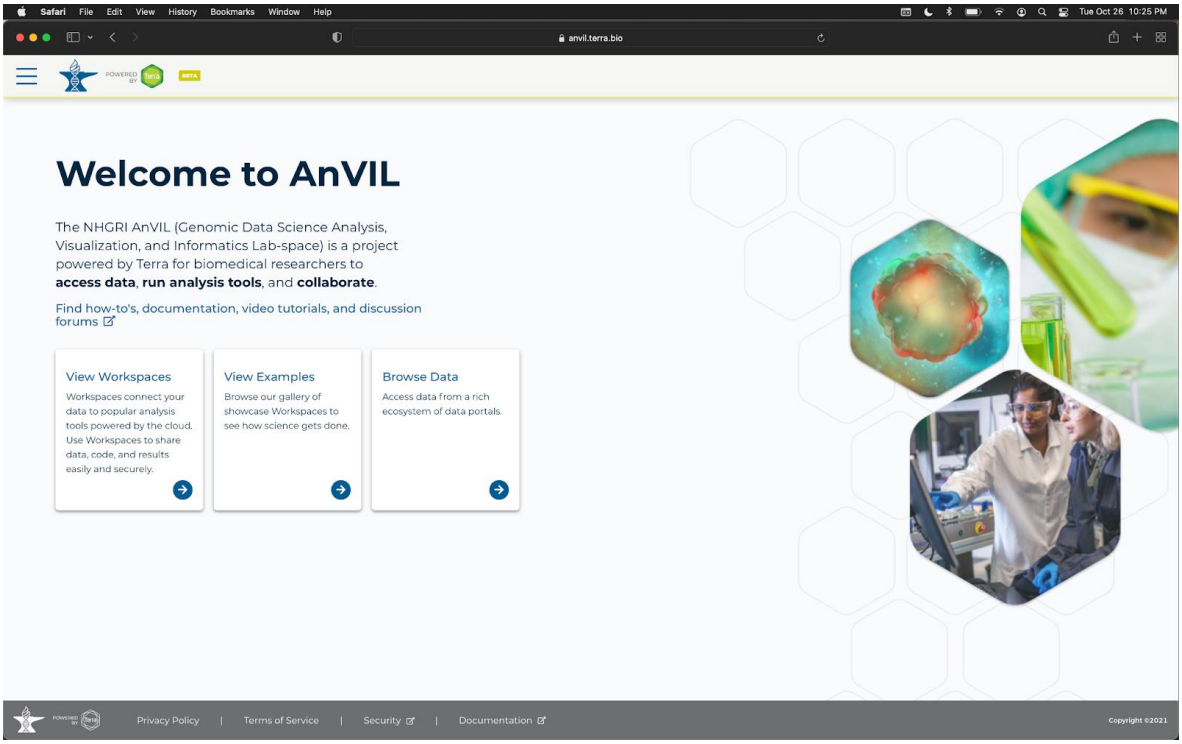
1. An Azure account and a subscription you can create resources in
2. Azure CLI
3. Create a resource group within your subscription

You can run Azure CLI or "az cli" commands to create, stop, restart or delete container instances running any bioconductor image - either official images by bioconductor or images available on Microsoft Container Registry. To get started, ensure you have an Azure account and a subscription or create a free account.

Follow this tutorial to get familiar with Azure Container Instances.

To run the bioconductor image hosted on Microsoft Container Registry or MCR, create a new resource group in your Azure subscription. Then run the following command using Azure CLI. You can customize any or all of the inputs. This command is adapted to run on an Ubuntu machine:

```
az container create \
        --resource-group resourceGroupName \
        --name mcr-bioconductor \
        --image mcr.microsoft.com/bioconductor/bioconductor_docker \
        --cpu 2 \
```



Microsoft Azure partnership with Bioconductor

Dr. Erdal Cosgun and his team at Microsoft Genomics

- Available now
  - Bioconductor docker images availability on MCR
  - Building binaries on Azure Kubernetes Service
  - Distribution of binaries
- Future
  - Annotation and Experiment Hub data on Azure

Acknowledgements

- Amazing Bioconductor Core Team
- Dr. Vince Carey
- Dr. Martin Morgan
- Dr. Rafael Irizarry

Questions

Email: **nitesh@ds.dfci.harvard.edu**

Questions to Bioconductor: **bioc-devel@r-project.org**

Join community-bioc on Slack!

SÖZLÜ BİLDİRİLER
ORAL PRESENTATIONS

S1

İndirgenmiş (Reduced) Rank Regresyon, Kısmi En Küçük Kareler ve Temel Bileşenler Faktör Analizi Yöntemlerinin Diyet Örüntüsünü Oluşturmada İstatistiksel Performanslarının Karşılaştırılması (Comparison of Statistical Performance of Reduced Rank Regression, Partial Least Squares, and Principal Components Factor Analysis Methods in Creating Diet Pattern)

Prof. Dr. Mehtap AKÇİL OK

Başkent Üniversitesi, Sağlık Bilimleri Fakültesi, Beslenme ve Diyetetik Bölümü, Ankara

Amaç: Hastalıkların oluşumunda ve gelişiminde tek bir besinin etkisi olmadığı, tüketilen tüm besin gruplarının ve alınan besin öğelerinin birlikte değerlendirilmesi epidemiyolojik beslenme araştırmalarında önemli yer almaktadır. Özellikle son yıllarda çok değişkenli veri yapısından daha az sayıda boyuta indirgeyerek diyet örüntüleri (dietary patterns) oluşturulmakta ve elde edilen skorlar ile hastalığa etki eden faktörler çok değişkenli istatistiksel analizlerle değerlendirilmektedir. Bu çalışmada, besin grupları ve öğeleri içeren gerçek bir veri kümesinde "İndirgenmiş (Reduced) Rank Regresyon Analizi (İRRA), Kısmi En Küçük Kareler Yöntemi (KEKKY) ve Temel Bileşenler Faktör Analizi (TBFA) Yöntemlerinin" diyet örüntüsünü oluşturmada istatistiksel performanslarının karşılaştırılmasını yapmak amaçlanmıştır.

Yöntem: Üç farklı boyut indirgeme yönteminin istatistiksel performanslarını karşılaştırmak için, 20-64 yaş arası 120 yetişkin bireyin besin gruplarının açıklayıcı ve besin öğelerinin cevap değişkenleri olarak alındığı çok değişkenli veri kümesi kullanılmıştır. İRRA, KEKKY ve TBFA yöntemlerinin matematiksel temelleri ve faktör çıkarma mantıkları benzer olmakla birlikte her bir yöntem cevap ve açıklayıcı değişkenlerin farklı kombinasyonlarını içeren kovaryans matrislerini kullanırlar. TBFA, açıklayıcı değişkenlerin kovaryans matrisini; İRRA, cevap değişkenlerinin arasındaki kovaryans matrisini ve KEKKY ise tüm açıklayıcı ve cevap değişkenlerinin arasındaki kovaryans matrislerini kullanarak faktör ayrıştırmasını ve skorlamasını yapmaktadır. TBFA tüm istatistiksel yazımlarda yer alırken, KEKKY ve İRRA için SAS ve R programına ve kodlarına gereksinim vardır. Genellikle yayınlanmış çalışmalar bu kodları ekler bölümünde paylaşmaktadır.

Bulgular: Diyetisyenlerin besin tüketim araştırmalarında yaygın olarak kullandıkları 70'den fazla gıda maddesi önerilen gruplara göre ayrılmış ve tüketilen besinlerden alınan makro ve mikro besin öğeleri BEBİS yazılım programında hesaplanarak veri kümesi oluşturulmuştur. KEKKY, İRRA ve TBFA yöntemleri uygulanmıştır. Boyut indirgeme açısından farklılık bulunmamıştır. Ancak seçilen dört besin öğesini açıklama varyanslarına göre performansları karşılaştırıldığında TBFA'nin besin öğelerindeki açıklama varyansı %25.6 ile %63.2; KEKKY'de %65.8 ile %89.2 ve İRRA de ise %78.9 ile %98.9 arasında değiştiği saptanmıştır. En yüksek açıklama oranı İRRA'da olup beslenmeye yönelik pek çok araştırmada da benzer sonuçlar bulunduğu görülmüştür.

Sonuç: Beslenme epidemiyolojisi alanında özellikle hastalık ile besin alımları arasındaki modellemelerde İndirgenmiş Rank Regresyon Analizi yöntemi ile diyet örüntüsü ve skorlarının kullanılması önerilmektedir.

Anahtar Sözcükler: İndirgenmiş Rank Regresyon Analizi, Kısmi En Küçük kareler Yöntemi, Temel Bileşenler Faktör Analizi, Diyet Örüntüsü

Kaynaklar

1. Duan MJ, Dekker LH, Carrero JJ and Navis G. Blood lipids-related dietary patterns derived from reduced rank regression are associated with incident type 2 diabets. *Clinical Nutrition*. 40: 4712-4713, 2021.
2. Gleason PM et all. Publishin Nutrition Research: A Review of Multivariate Techniques-Part3: Data Reduction Methods. *J Acad Nutr diet.* 115:1072-1082, 2015.
3. Sangsefidi ZS et all. Major dietary patterns and differentiated throid cancer. *Clinical Nutrition ESPEN. 33:195-201. 2019.*
4. Wagner S et all. Association of dietary patterns derived using reduced rank regression with subclinical cardiovascular damage according to generation and sex in the STANISLAS Cohort. *J Am Heart Assoc*. 9:1-13, 2020.
5. Davies PT and TSO M. Procedures for reduced rank regression. *Appl.Statist.* 31(3): 244-255, 1982.

S2

Comparison of Three Methods Used in Analysis of Longitudinal Data on Health: ANOVA, GEE, QLS

Erdoğan ASAR[1], Erdem KARABULUT[2]

[1] Turkish Statistical Institute, Statistical Consultancy, Ankara, Turkey
[2] Hacettepe University, Medicine Faculty, Department of Biostatistics, Ankara, Turkey

Introduction: Researchers can make multiple measurements (repeated measurements) on the same observation unit in order to reduce the errors caused by the observation units and to see the changes in the observation units over time. Such data based on repeated measurements are called longitudinal data. Among the methods used in the analysis of such data are linear models such as Analysis of Variance (ANOVA), as well as Generalized Estimating Equations (GEE) developed from Generalized Linear Models (GLM) and Quasi-Least Squares Regression (QLS) developed from GEE. In this study, it was aimed to analyze the longitudinal data in the field of health with ANOVA, GEE and QLS methods and compare the results.

Method: Analyzes were made on a longitudinal data set related to an experimental study in the field of health. The data set consists of 7 quantitative measurements in terms of the dependent variable studied for the two groups, and each group consists of 10 observation units. Quantitative measurements on two groups; It was done at unequal time intervals, 7, 14, 28, 60, 120, 180 and 220 days. Another remarkable feature in the obtained data set is that there are missing measurements in some observations due to the inability to measure for any reason at some measurement times. The analysis of whether there is a significant difference between the two groups in terms of dependent variable was made using ANOVA, GEE and QLS methods, which can be used for such data sets. The results were compared and comments were made. Regression models in QLS and GEE were constructed using Exchangeable, Tri-diagonal, The First Order Autoregressive (AR-1) and Markov working correlation structures.

Findings: In the analyzes performed with ANOVA, GEE and QLS, it was seen that the time effect on the dependent variable created a significant difference ($p<0.05$), while the interaction effect (the effect of time and group effect together on dependent variable) was not significant ($p>0.05$). It was concluded that the group effect on the dependent variable was not significant in the ANOVA, that is, there was no statistical difference between the groups according to the dependent variable ($p>0.05$). On the contrary, the group effect in GEE and QLS was significant for all models except the model created with AR-1 correlation structure ($p<0.05$).

Result: It was found in all analysis methods that measurement times had an effect on the dependent variable. In the group effect on the dependent variable, the results of ANOVA and GEE-QLS are inconsistent. While the group effect was not significant in ANOVA, it was found to be significant in the other two methods in general. In addition, while the group effect was not found significant in both GEE and QLS according to the model established according to the AR-1 working correlation structure, which takes into account the measurement orders, the group effect was found to be significant compared to the model established with the Markov correlation structure in which the actual measurement times were used in QLS. It is concluded that it would be appropriate to use the Markov working correlation structure in cases where repeated measurements are not made at equal time intervals.

Keywords: Longitudinal data, analysis of variance, quasi-least squares regression, generalized estimating equations, markov correlation structure.

References

1. Alpar R. Çok Değişkenli İstatistiksel Yöntemler. 5.Baskı. Ankara: Detay Yayıncılık; 2017.
2. Barnett AG, Koper N, Dobson AJ, Schmiegelow F, Manseau M. Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology. Methods in Ecology and Evolution. 2010; 1: 15-24.
3. 11. Chaganty NR. An alternative approach to the analysis of longitudinal data via generalized estimating equations. Journal of Statistical Planning and Inference. 1997; 63: 39-54.
4. 13. Chaganty NR, Shults J. On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. Journal of Statistical Planning and Inference. 1999; 76: 145-161.

5. Davis CS. Statistical Methods fort he Analysis of Repeated Measurements. New York: Springer-Verlag; 2002.
6. Hardin JW, Hilbe JM. Generalized Estimating Equations. Second Edition. New York: Taylor&Francis Group; 2013.
7. Hedeker D, Gibbons RD. Longitudinal Data Analysis. New Jersey: John Wiley & Sons, Inc.; 2006.
8. Kim H, Shults J. %QLS SAS Macro: A SAS Macro for Analysis of Correlated Data Using Quasi-Least Squares. Journal of Statistical Software. 2010; 35(2): 1-22.
9. Pan W, Connett JE. Selecting The Working Correlation Structure In Generalized Estimating Equations With Application To The Lung Health Study. Statistica Sinica. 2002; 12(2): 475-490.
10. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika. 1990;77(3): 485-497.
11. Shi G, Chaganty R. Application of Quasi-Least Squares to Analyse Replicated Autoregressive Time Series Regression Models. Journal of Applied Statistics. 2004; 31(10): 1147-1156.
12. Shults J, Chaganty NR. Analysis of Serially Correlated Data Using Quasi-Least Squares. Biometrics. 1998; 54(4): 1622-1630.
13. Shults J, Hilbe JM. Quasi-Least Squares Regression. New York: Taylor&Francis Group; 2014.
14. Shults J, Morrow AL. Use of Quasi-Least Squares to Adjust for Two Levels of Correlation. Biometrics. 2002; 58(3): 521-530.
15. Shults J, Ratcliffe SJ, Leonard M. Improved generalized estimating equation analysis via xtqls for implementation of quasi-least squares in Stata. Upenn Biostatistics Working Papers. 2006; 13: 1-21.
16. Smith T, Smith B. Proc Genmod with GEE to Analyze Correlated Outcomes Data Using SAS [Internet]. 2018 [Erişim Tarihi 29 Mart 2019]. Erişim adresi: https://www.lexjansen.com/wuss/2006/tutorials/TUT-Smith.pdf.
17- StataCorp. 2015. Stata Statistical Software: Release 14.1. Texas: StataCorp LP.
18. Wang M. Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. Advances in Statistics. 2014; 303728: 1-11.
19. Xie J, Shults J. Implementation of quasi-least squares With the R package qlspack. UPenn Biostatistics Working Papers. 2009; 32: 1-19.
20. Ziegler A. Generalized Estimating Equations. New York: Springer; 2011.

S3

Çok Kategorili Hastalık Durumlarında Tanısal Modele Yeni Bir Belirteç Eklenmesinin Tanı Performansındaki Değişime Etkisinin İncelenmesi

Hande ŞENOL[1], Ergun KARAAĞAOĞLU[2]

[1] Pamukkale Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Denizli
[2] Hacettepe Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara

Amaç: Sağlık alanında, hasta ve sağlıklı bireylerin belirlenmesinde, tanı kestirimi modellerinden elde edilen risklerin doğru değerlendirilmesi önemlidir. İki kategorili sınıflandırmalarda en sık kullanılan yöntem olan ROC analizi yönteminden yola çıkarak çok kategorili sınıflandırma problemleri için HUM (Hypervolume Under The ROC Manifold) ve CCP - Doğru Sınıflandırma Olasılığı (Correct Classification Probability) yöntemleri geliştirilmiştir. Ancak, bu yöntemlerin var olan bir modele yeni bir belirteç eklendiğinde meydana gelen performans artışını ölçümlemede başarılı olmadığı düşünülmektedir. Bu nedenle, yeni bir belirteçin modele eklendiğinde oluşturulan yeni sınıflama sonucunun performansını gösteren iki performans ölçüsü geliştirilmiştir. Bunlar; NRI (Net Reclassification Improvement) ve IDI (Integrated Discrimination Improvement)'dır.

Yöntem: Bu çalışmada, çok kategorili sınıflandırma modellerinde kullanılan performans ölçüleri HUM ve CCP'nin, model performansına etki eden belirteçlerin etkisini ölçen NRI ve IDI ile ilişkilerinin incelenmesi amaçlanmıştır. Bağımlı değişkenin farklı tipte olduğu durumlar göz önüne alınarak (ordinal ve nominal), belirteçlerin modele eklenme sırasının model performansına etkisi araştırılmıştır. İncelemeler, gerçek veri setleri ve bir benzetim çalışması ile yapılmıştır. Benzetim çalışmasında, hem belirteçlerin kendi arasında hem de belirteçler ve bağımlı değişken arasında olmak üzere; pozitif, negatif, düşük, orta ve kuvvetli seviyelerde ilişki yapıları ve ordinal yapıda bir bağımlı değişken tasarlanmıştır.

Bulgular: Sonuçlarda, negatif ilişki yapısında benzetim yapılan verilerden elde edilen modellerin NRI ve IDI değerlerinin klasik yöntemler olan HUM ve CCP değerleri ile daha yüksek düzeyde ve pozitif yönde ilişkilere sahip olduğu görülmüştür. Gerçek veri setleri ile yapılan iki ayrı uygulamada ise tanısal modele eklenen belirtecin sırasının önemini ölçebilmek için NRI ve IDI sonuçları incelenmiş ve tahmin başarısı en yüksek olan model oluşturulurken, belirteçlerin modele eklenme sırasının model performansında meydana getirdiği artış ve azalışlar incelenebilmiştir. Elde edilen NRI ve IDI sonuçları ile en az sayıda belirteç modele eklenerek en yüksek tahmin başarısına sahip olan modele ulaşılmıştır.

Sonuç: NRI ve IDI yöntemleri, HUM ve CCP yöntemlerinin aksine belirtecin modele eklenme sırasından etkilenmektedir. Bu yöntemler kullanılarak, doğru sırada eklenen daha az sayıda belirteç ile doğru sınıflandırma performansının arttırılabileceği sonucuna ulaşılmıştır. Bu yöntemler, gerek bağımlı değişkenin gerekse modelleme için incelenecek olan belirteçlerin, her değişken tipinde (sayısal, nominal ve ordinal) kullanılabilmesi ve tüm regresyon modelleri ile uygulanabilmesi açısından da tercih edilen yöntemlerdir. Ayrıca, bu yöntemler parametrik test varsayımları gerektirmediğinden, literatürde bulunan birçok farklı yönteme göre kullanım açısından oldukça avantajlıdır.

Anahtar Sözcükler: Çok kategorili sınıflandırma, HUM, CCP, NRI, IDI

S4

Yapısal Eşitlik Modeli ile Uzunlamasına Veri Analizi: Örtük Büyüme Modelleri ve Gerçek Veri Seti Üzerinde Bir Uygulama

Funda Seher ÖZALP ATEŞ[1], Derya GÖKMEN[2], İpek GÖNÜLLÜ[3]

[1] *Manisa Celal Bayar Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Manisa*
[2] *Ankara Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara*
[3] *Ankara Üniversitesi Tıp Fakültesi, Tıp Eğitimi ve Bilişimi Anabilim Dalı, Ankara*

Amaç: Bireylerin ilgilenilen özellikleri açısından zamana bağlı değişimlerinin incelenmesini amaçlayan araştırmalarda, bireylerden farklı zaman noktalarında elde edilen tekrarlı ölçümlerin değerlendirilmesinde kullanılan yöntemlerden birisi de Örtük Büyüme Modelleri (ÖBM)'dir. ÖBM ile ilgilenilen özellik bakımından hem birey içi hem bireyler arası farklılıklar değerlendirilebilir. Bu çalışmanın amacı tek değişkenli örtük büyüme modelini incelemek ve gerçek bir veri seti üzerinde uygulayarak sonuçlarını irdelemektir.

Yöntem: ÖBM, iki düzeyli bir analizdir. Düzey-1'de birey içi değişim, Düzey-2'de ise birey içi değişimdeki bireyler arası farklılıklar değerlendirilir. ÖBM'de gözlenen değişkenin tekrarlı ölçümleri modelde "gösterge" değişken olarak yer alır. Büyüme eğrilerini tanımlamak amacıyla gözlenen değişkenler arasındaki ilişkileri açıklamak için örtük değişkenler kullanılır. ÖBM, sabit (π_0) ve eğim (π_1) olmak üzere iki örtük faktör içerir. Sabit ve eğim faktörleri için ortalamalar, varyanslar ve iki faktör arasındaki kovaryans değeri tahmin edilir. ÖBM'de tekrarlı ölçülen değişken kategorik olduğunda kullanılan parametre kestirim yöntemlerinden biri olan örtük yanıt değişkeni (ÖYD) dönüşümünde, kategorik yanıtlar, modele ek örtük değişkenler eklenerek normal dağılım gösteren sürekli değişkenlere (Y*) dönüştürülür. Sabit faktörünün ortalama değeri (η_0), örtük yanıt değişkeni Y*'ın ortalama başlangıç düzeyini ifade eder ve modelin tanımlanabilir olması için 0'a eşitlenir. Eğim faktörünün ortalama değeri ise, zamandaki bir birimlik artışa karşılık gelen Y*'daki değişim oranını gösterir.

Bu çalışmada kullanılan veriler, Ankara Üniversitesi Tıp Fakültesi'nde, 2012-2013 Eğitim-Öğretim Yılı'nda eğitime başlayan öğrencilere belirli dönemlerde olmak üzere toplamda 7 defa 19 soruluk "Jefferson Doktor Empati Ölçeği Öğrenci Versiyonu" uygulanarak elde edilmiştir.

Bulgular: Verilerin ÖBM ile analize uygun olup olmadığının değerlendirilmesi amacıyla öncelikle tekrarlı ölçümler arasındaki korelasyonlar incelenmiştir ve verilerin ÖBM ile analize uygun olduğu belirlenmiştir. Bileşik puanlarla doğrusal, karesel ve parçalı ÖBM analiz edilmiştir ve tümel olarak uyum indisleri değerlendirildiğinde parçalı ÖBM (Yaklaşımın Hata Kareler Ortalaması Karekökü: 0,048, Karşılaştırmalı Uyum İndeksi: 0,963, Tucker Lewis İndeksi: 0,969 ve $\chi 2$/sd: 1,724) en uygun model olarak elde edilmiştir. Parçalı ÖBM sonuçlarına göre, öğrencilerin klinik öncesi dönemde (1. eğim değeri -0,052 (p=0,007)) empatik olma eğilimlerinde zaman boyunca düşüş varken; klinik dönemde (2. eğim değeri 0,044 (p=0,086)) benzer oranda daha empatik olma eğiliminde oldukları belirlenmiştir. Ancak bu eğilimler öğrenciler arasında farklılık göstermemektedir (1. eğimin varyans değeri 0,002 (p=0,840), 2. eğimin varyans değeri 0,013 (p=0,260)).

Sonuç: Sağlık alanında bireylerin ilgilenilen bir özellik açısından zamana bağlı değişimlerinin incelenmesi amacıyla elde edilen uzunlamasına verilerin değerlendirilmesinde kullanılan yöntemlerden biri ÖBM'dir. ÖBM; birey içi değişimdeki bireyler arası farklılıkları değerlendirebilmesi, karmaşık tasarım modellerde uygulanabilmesi ve özellikle de hataların modele dahil edilebilmesi bakımından diğer yöntemlere göre bazı avantajlar sağlamaktadır.

Anahtar Sözcükler: Kategorik Veri, Örtük Büyüme Modelleri, Uzunlamasına Veri Analizi

S5

Analysis of the Relationships between Life Satisfactions, Stress Levels, Eating Attitudes and Sleep Problems of People

Muharrem Gürleyen GÖK[1], Erdem KARABULUT[1]

[1] Hacettepe University Faculty of Medicine, Department of Biostatistics, Ankara

Abstract

Introduction: Persons interact with many different social and economic factors in their life cycle from birth to death, such as the education they receive, the work they do, their marriage, family life, health status and income. In other words, they are both physiologically and psychologically affected by these elements and affect them with their attitudes and behaviours. In this context, the aim of the study is to examine the interactions between life satisfaction, stress levels, sleep problems and eating attitudes of the individuals.

Method: In the study, the results of an online survey conducted to a public institution personnel were analyzed by using Nonlinear Canonical Correlation Analysis (OVERALS).

In addition to the questions such as gender, age, weight, eating habits (fast, excessive, irregular eating), satisfaction with weight-related appearance; the Turkish versions of the scales named Three Factor Eating Questionnaire (TFEQ), Eating Attitudes Test (EAT), Satisfaction with Life Scale (SWLS), Perceived Stress Scale (PSS) and Epworth Sleepiness Scale (ESS) were also included in the survey. The TFEQ measures 3 aspects of eating behaviour of the individuals: Cognitive restraint (degree of consciously restricting their eating), uncontrolled eating, and emotional eating (degree of eating when they are emotional). EAT specifies the possible disorders in eating behaviour, EUS determines whether there is a problem of daytime sleepiness, SWLS evaluates life satisfaction, and PSS measures the level of stress.

Results: Out of the total 175 participants, 49% were men and 51% were women. In the age distribution, the highest proportion belongs to the employees in the 40-49 age group with 33%. This is followed by participants in the 50-59 age group with 28%, 30-39 age group with 21%, and 20-29 age group with 18%. Considering the sub-dimensions of the scales used, multiple analyzes were carried out using different variable sets, as there was a large data set in terms of the number and variety of the derived variables. In the first analysis using the results obtained from the above-mentioned scales, it was determined that 68% of the possible maximum correlation between the two variable sets was related to the current model (fit value=1.367).

Conclusion: As a result of the analyzes carried out with different sets of variables, for the participants of the survey it has been determined that: Persons who fail in emotional and uncontrolled eating have high stress and daytime sleepiness problems. Persons with average level of life satisfaction do not have daytime sleepiness problem and their eating attitudes are normal. Women are successful in cognitive restraint and uncontrolled eating, but unsuccessful in emotional eating. Men are successful in emotional eating but unsuccessful in cognitive restraint and uncontrolled eating, and they do not care about their weight-related appearance. Overweight and obese people are dissatisfied with their weight-related appearance and they fail in uncontrolled eating. People with low perceived stress levels do not experience disordered eating and waking up tired. People with daytime sleepiness and waking up tired often have an excessive and irregular eating attitude.

Keywords: OVERALS, Eating Attitudes, Life Satisfaction, Stress

References

1. Alpar, R. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 4. Baskı, Detay Yayıncılık, Ankara, 2013.
2. Altaş, D. Giray, S. Dünyadaki En Önemli Sorun Algısının Optimal Ölçeklemeli Çok Değişkenli İstatistiksel Yöntemler İle İncelenmesi. Öneri Dergisi - Cilt 10, Sayı 39 (2013). http://e-dergi.marmara.edu.tr/maruoneri/article/view/1012000313
3. Dağlı, A. Baysal, N. Yaşam Doyumu Ölçeğinin Türkçe'ye Uyarlanması: Geçerlik ve Güvenirlik Çalışması. Elektronik Sosyal Bilimler Dergisi, Güz-2016 Cilt:15 Sayı:59 (1250-1262) https://toad.edam.com.tr/sites/default/files/pdf/yasam-doyumu-olcegi-toad.pdf

4. Eskin, M. Harlak, H. vd. Algılanan Stres Ölceğinin Türkceye Uyarlanması: Güvenirlik ve Geçerlik Analizi. New/Yeni Symposium Journal, Ekim 2013, 132 | Cilt 51 | Sayı 3 http://yenisymposium.com/Pdf/EN-YeniSempozum-c1d2631c.PDF
5. Fleurbaix Laventie Ville Sante (FLVS) Study Group. The Three-Factor Eating Questionnaire-R18 Is Able to Distinguish among Different Eating Patterns in a General Population. The Journal of Nutrition, Volume 134, Issue 9, 1 September 2004, Pages 2372–2380. https://academic.oup.com/jn/article/134/9/2372/4688768
6. Garner, D. Garfinkel, P. The Eating Attitudes Test: An Index of the Symptoms of Anorexia Nervosa, Psychological Medicine, 1979, 9, 273-279. https://www.researchgate.net/publication/22670350_The_Eating_Attitudes_Test_An_Index_of_the_Symptoms_of_Anorexia_Nervosa
7. Garson, D. GLM Multivariate, Manova & Canonical Correlation. Statistical Associates Publishing Blue Book Series, ISBN:978-1-62638-034-9, 2015.
8. Gifi, A. Algorithm Descriptions for ANACOR, HOMALS, PRINCALS, and OVERALS. Research Report RR-89-01, Department of Data Theory FSW/RUL, Leiden, The Netherlands, 1989. http://gifi.stat.ucla.edu/janspubs/Author/GIFI-A.html
9. Gifi, A. Nonlinear Multivariate Analysis. Wiley, New York, 1990.
10. İzci, B. Ardıç, S. vd. Reliability and validity studies of the Turkish version of the Epworth Sleepiness Scale, Sleep Breath (2008) 12:161–168 DOI 10.1007/s11325-007-0145-7 https://toad.edam.com.tr/sites/default/files/pdf/epworth-sleepiness-scale-toad.pdf
11. Kıraç, D. Kaspar, E.Ç. vd. Obeziteyle İlişkili Beslenme Alışkanlıklarının Araştırılmasında Yeni Bir Yöntem "Üç Faktörlü Beslenme Anketi." MÜSBED 2015;5(3):162-169 http://dergipark.gov.tr/download/article-file/165495
12. Meulman, JJ. Heiser, WJ. IBM SPSS Categories 20. SPSS Inc., USA, 2011
13. Oral, N. HİSLİ ŞAHİN, N. Yeme Tutum Bozukluğunun Kişilerarası Şemalar, Bağlanma, Kişilerarası İlişki Tarzları ve Öfke ile İlişkisi. Türk Psikoloji Dergisi, Aralık 2008, 23 (62), 37-48. https://www.researchgate.net/publication/227854860_Yeme_Tutum_Bozuklugunun_Kisilerarasi_Semalar_Baglanma_Kisilerarasi_Iliski_Tarzlari_ve_Ofke_ile_Iliskisi
14. Savaşır, E.N. Yeme Tutum Testi: Anoreksia nevroza belirtileri indeksi. Türk Psikoloji Dergisi 1989; 23:132-6. https://toad.edam.com.tr/sites/default/files/pdf/yeme-tutum-testi-anoreksiya-nervoza-belirtileri-indeksi-toad.pdf
15. Thanoon, Y. Robiah, A. Saffari, S. Generalized Nonlinear Canonical Correlation Analysis With Ordered Categorical And Dichotomous Data. 2014. http://www.jurnalteknologi.utm.my/index.php/jurnalteknologi/article/view/3602
16. TOAD Türkiye Ölçme Araçları Dizini, Epworth Uykululuk Ölçeği. https://toad.edam.com.tr/olcek/epworth-uykululuk-olcegi
17. Van der Burg, E. De Leeuw, J. and Dijksterhuis, G. Non-Linear Canonical Correlation With K Sets of Variables. Computational Statistics and Data Analysis, 18: 141 - 163, 1994. http://gifi.stat.ucla.edu/janspubs/Author/VAN-DER-BURG-E.html
18. Van de Geer, J.P. Algebra and Geometry of Overals. 1987. http://www.datatheory.nl/pdfs/87/87_13.pdf

S6

Sağkalım Analizinde Derin Öğrenme

İrem KAR[1], Erdal COŞGUN[2], Atilla Halil ELHAN[1]

[1]*Ankara Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara*
[2] *GenomicsTeam, Microsoft Research&AI, Redmond, WA, USA*

Amaç: Bu çalışmada, sağkalım analizine derin öğrenme tabanlı bir yaklaşım olan DeepSurv uygulanması ve sonuçların makine öğrenmesi yöntemlerinden RSF ve klasik istatistiksel yöntemlerden Cox regresyon ile karşılaştırılması amaçlanmıştır.

Yöntem: Burada, türetilmiş veri setleri kullanılarak bir benzetim çalışması gerçekleştirilmiştir. Bu kapsamda örneklem büyüklüğü ile sansür oranının çeşitli düzeylerinin ele alındığı ve bağımsız değişkenlere ilişkin katsayıların tekdüze dağılımdan türetildiği çeşitli senaryolar oluşturulmuştur. Oluşturulan her senaryo ise 100 kez tekrarlanmıştır. Belirlenen senaryolara ek olarak, bağımsız değişken katsayılarına ilişkin ön belirlemenin yapılmadığı, R programlama dilindeki "coxed" paketindeki mevcut haliyle bırakıldığı ayrı senaryolar da oluşturulmuştur. Yöntemlerin karşılaştırılmasında Harrell'in C-indeks değeri kullanılmıştır. Verilerin türetilmesi, RSF ve Cox regresyon yöntemlerinin uygulanması R programlama dili (ver. 3.6.3) ile gerçekleştirilmiştir. Derin öğrenmeye ilişkin analizler ise Python programlama dilinde gerçekleştirilmiştir.

Bulgular: Yapılan benzetim çalışması sonucunda, bağımsız değişken katsayılarının tekdüze dağılımdan türetildiği senaryolarda tüm yöntemlerin tahminleme başarısı yüksek bulunmuştur. Örneklem büyüklüğü arttıkça yöntemlerin performansı artmıştır. Sansür oranının artması DeepSurv ve RSF yöntemlerinin tahminleme başarısını azaltmıştır. Bağımsız değişken katsayılarının paketteki mevcut haliyle kullanılıp normal dağılımdan türetildiği durumda ise üç yöntem sonucunda da 0,50-0,60 civarında bir C-indeks değeri elde edilmiştir.

Sonuç: Senaryolardaki farklılaşmanın etkisi Cox regresyonda gözlemlenemezken, RSF ve DeepSurv yöntemlerinde parametrelerdeki değişimin etkisi açık bir şekilde ortaya konmuştur. Sonuç olarak, derin öğrenme oldukça hızlı gelişen bir alandır. DeepSurv algoritmasında çok sayıda hiperparametre kullanılmasına rağmen, mimarisi anlaşılır ve basittir. Hiperparametre sayısının çok olması, model başarısını arttırmak için avantajlıdır.

Anahtar Sözcükler: Cox regresyon, DeepSurv, Harrell'in C-indeksi, Random Survival Forest, Sağkalım Analizi

Kaynaklar

1.Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology, 18(1), 1-12.
2. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The annals of applied statistics, 2(3), 841-860.
3. Kropko, J., & Jeffrey, J. H. (2019). coxed: An R Package for Computing Duration-Based Quantities from the Cox Proportional Hazards Model. R J., 11(2), 38.
4. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. Jama, 247(18), 2543-2546.

S7

Kalp Yetmezliği Hastalarında Sağkalım Tahmini İçin Makine Öğrenmesi Modellerinin Performanslarının Karşılaştırılması ve Sağkalım ile İlişkili En Önemli Faktörlerin Belirlenmesi

Rüstem YILMAZ[1], **Fatma Hilal YAĞIN[2]**

[1]Samsun Gazi Devlet Hastanesi, Kardiyoloji Bölümü, İlkadım, Samsun, Türkiye
[2]İnönü Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya, Türkiye.

Giriş: Bu çalışmada kalp yetmezliği olan hastaların sağkalım durumunu tahmin etmek için makine öğrenmesi yöntemlerinin performanslarının karşılaştırılması ve kalp yetmezliği hastalarının sağkalımını en çok etkileyen faktörlerin belirlenmesi amaçlanmıştır.

Yöntem: 299 kalp yetmezliği hastasının tıbbi kayıtlarını içeren bir veri seti analiz edilmiştir. Veri setindeki sınıf dengesizliği problemini (ölen: 96 (% 32.1), yaşayan: 203 (% 67.9)) giderebilmek için oversampling yöntemi kullanılmıştır. Lojistik Regresyon, Rasgele Orman ve Destek Vektör Makinesi algoritmaları ile 3 ayrı model oluşturulmuştur. Hiperparametre optimizasyonu için 10 katlı 5 tekrarlı çapraz geçerlilik ile ızgara arama yöntemi kullanılmıştır. Oluşturulan modellerin performansını değerlendirmek için doğruluk, duyarlılık, seçicilik ve F1 skor değerleri hesaplanmıştır. Ek olarak kalp yetmezliği sağkalımını en çok etkileyen faktörleri bulmak için değişken önemlilikleri incelenmiştir. Analizler Python programında gerçekleştirilmiştir.

Bulgular: Sonuçlar incelendiğinde, Rasgele Orman modeli kalp yetmezliği sağkalım tahmini için en iyi performansı gösterdi (Doğruluk: % 90.2, Duyarlılık: % 94.6, Seçicilik: % 86.7, F1 Skor: % 89.7). Serum kreatinin ve ejeksiyon fraksiyonu değişkenleri kalp yetmezliği olan hastaların sağkalımını tahmin etmek için en önemli iki faktör olarak bulunmuştur.

Sonuç: Elde edilen sonuçlar, bir kalp yetmezliği hastasının hayatta kalıp kalamayacağını tahmin ederken sağlık çalışanları için yeni bir destekleyici araç haline gelerek klinik uygulamayı etkileme potansiyeline sahip olabilir. Kalp yetmezliği olan bir hastanın hayatta kalıp kalamayacağını tahmin etmenin yanı sıra bu hastalarda ilk olarak serum kreatinin ve ejeksiyon fraksiyonuna odaklanabilir.

Anahtar Sözcükler: Kalp Yetmezliği, Sınıflama, Parametre Optimizasyonu, Değişken Önemliliği.

S8

Detection of Anaemia and Anaemia-Related Factors in Pregnant Women by Supervised Machine Learning Methods

Rüveyda YILDIRIM[1], Mehmet Onur KAYA[1], Burkay YAKAR[2], Bilal ALATAŞ[3]

E-mail: ruveyda23@hotmail.com

[1] Fırat University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Elazığ
[2] Fırat University Faculty of Medicine, Department of Family Practice, Elazığ
[3] Fırat University Faculty of Engineering, Department of Software Engineering, Elazığ

Abstract:

Introduction: In the study, it was aimed to establish a clinical decision support system by detecting pregnancy anemia and related factors with supervised machine learning methods in the pregnant population living in the city center of Elazığ.

Method: Our patients were randomly selected from 3228 general pregnant population registered in family health centers in the first half of 2019, according to population size, as a sample of roughly 489 pregnant women to represent all pregnant women residing in the city center of Elazig. Using the Java-based open source Weka software and the study's data, supervised machine learning approaches were used to develop classification models that can diagnose anemia. The anemia detection system, which was created with rule-based Jrip, OneR, and PART algorithms, was compared in test data according to different success metrics such as Accuracy, Precision, and AUC (Area Under The Receiver Operating Characteristic Curve). For the test data, the 5-fold cross-validation method of the training set was used.

Findings: In the study, success criteria were obtained according to the 5-fold cross-validation method of three algorithms. The Accuracy, Precision, and AUC values of the Jrip algorithm were found as $Jrip_{ACC}$=0.919, $Jrip_{PRE}$=0.923, $Jrip_{AUC}$=0.947 respectively. The Accuracy, Precision, and AUC values of the OneR algorithm were found as $OneR_{ACC}$=0.854, $OneR_{PRE}$=0.850, $OneR_{AUC}$=0.795 respectively. The Accuracy, Precision, and AUC values of the PART algorithm were found as $PART_{ACC}$=0.919, $PART_{PRE}$=0.922, $PART_{AUC}$=0.942 respectively.

Result: According to the 5-fold cross-validity test type of rule-based algorithms, the Jrip method produces the greatest results in terms of the classification models' Accuracy, Precision, and AUC success metrics, while the OneR algorithm produces the worst results.

Key Words: Machine Learning, Classification, Anaemia

S9

Evaluation of Traditional Machine Learning and Mixed Effect Machine Learning Model Performances by Simulation Study

Ebru TURGAL[1], Beyza DOĞANAY ERDOĞAN[2]

[1,2] *Ankara University, Faculty of Medicine, Department of Biostatistics, Ankara.*
e-posta: ebruturgal@gmail.com

1. Objective

Development of machine learning has increased rapidly in recent years and many algorithms have been developed in this regard. So far, only one study has been found examining the performance of mixed-effects machine learning models in grouped data [42].

In this study, the efficiency of mixed-effects models (LME), which is one of the standard models used in classification of grouped data and mixed-effects machine learning models, which is a new approach, will be compared under different scenarios.

2. Method

2.1 Community Methods

2.1.1.Bagging

This algorithm is the abbreviation of bootstrap and aggregating terms and was put forward by Breiman in 1996 [62]. The purpose of the bagging algorithm is to generate new datasets randomly by using the training dataset, thus make differences between clusters and increasing the classification success. In addition, it aims to retrain the classification model by deriving new sets from the training dataset with the random selection method by with replacement. The first step in the bagging method is to divide the dataset into a training set and a test set. From the training set containing n samples, one or more new training sets with n samples are produced by with replacement by random selection method. Each classifier in the ensemble formed by the bagging method is trained with a training set containing different samples. The result of each classifier is combined with majority voting [9,62,63].

2.1.2. Boosting

Boosting algorithm was developed by [49] in 1999. Boosting algorithm is a sequential method based on slow learning and aims to learn from error. These algorithms combine several low-precision models to create high-precision models. The aim is to try to obtain a strong model by combining the models obtained in each iteration within the framework of certain rules [62,63].

2.1.3. Gaussian Process

Gaussian processes [60] are flexible non-parametric function models that achieve state-of-the-art estimation accuracy and allow making probabilistic estimations [20].

The Gaussian process is a non-parametric Bayesian approach to regression used to approximate functions [5]. The Bayesian approach determines a probability distribution over possible functions. [29]. The algorithm for this process is a powerful algorithm for modeling nonlinear relationships between the pairs of random variables. It defines a distribution over functions that can be applied to demonstrate uncertainty about the actual functional relationship [60].

Gaussian process are used in areas such nonparametric regression, modeling of time series [53], spatial [6] and spatio-temporal [10] data, emulation of large computer experiments [28] optimization of expensive black box functions [27] and parameter tuning in machine learning models [55].

The Bayesian approach determines a probability distribution over all possible values. In this approach, parameters are behaved as previously distributed random variables. Other supervised learning methods also learn exact values for each parameter. Gaussian process regression calculates the probability distribution over all acceptable functions that fit the data. The a priori point is determined, the posterior point is calculated using the training data and

compared with the estimated posterior at the respective points. In this method, result of $\sum(y - \hat{y})^2$ equation should be minimal [5].

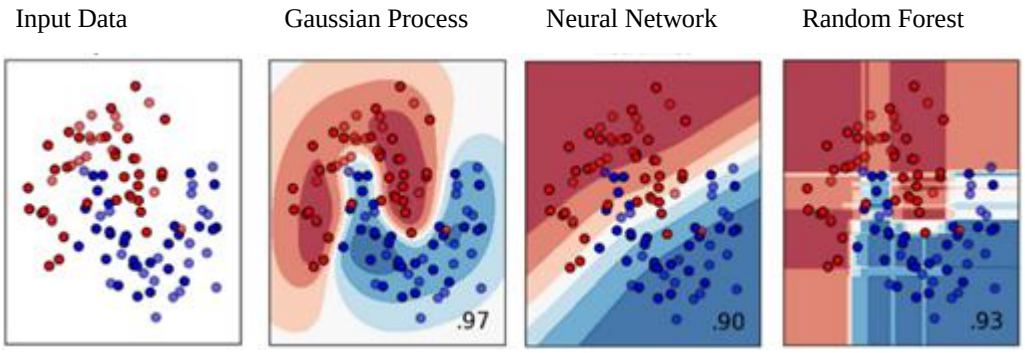


Figure 1: Classifier comparison outputs

Figure 1 shows the classification functions learned by the different methods used to separate the blue and red dots. Neural network and random forest, which are widely used and powerful methods, produce predictions away from the training data. The Gaussian process obtains the model output with greater accuracy, which is particularly important in authentication and security-critical uses [29].

In addition, the gaussian process is widely used in motion planning, positioning, localization and mapping in the field of robotics [24,62].

2.1.4. Gaussian Process Boosting

The model is trained using the GPBoost (Gaussian Process Boosting) algorithm. This takes place by training the covariance parameters of random effects and the mean F(X) function with a tree ensemble model. The algorithm used is an boosting algorithm that iteratively learns the covariance parameters and adds a tree to the tree ensemble using a gradient or Newton acceleration step. Covariance parameters can be learned using (Nesterov acceleration) gradient descent or Fisher scoring [54].

For finite samples, boosting tends to overfit and to avoid this, early stopping should be applied, especially in regression tasks.

2.1.5. Mixed Effect Model with Gaussian Process

In mixed-effects models of the Gaussian process, the first moment is usually assumed to be either zero or a linear function of the covariates. The residual structured variation is then modeled using the zero-mean Gaussian process and/or the grouped random effects model. However, both the zero mean and linearity assumption are often unrealistic and higher prediction accuracy can be achieved by relaxing these assumptions [54]. Also, if the mean function of a Gaussian process model is misidentified, spurious second-order non-stationarity can occur because the covariance function of such not well-defined model is equal to the true covariance function plus the square of the error of the mean function [2,17].

In state-of-the-art supervised machine learning algorithms, particularly with boosting, a flexible and potentially complex function is assumed to associate a set of predictive variables with a response variable. Based on the predictive variables, the response variable is assumed to be independent between observations. This means that potential residual correlation, where, correlation not explained by the regression function, is ignored [54]. Briefly, this approach proposes to model the mean function and the predicted parameters of the covariance structure of random effects with the mean function with a community of basic learners, such as regression trees that learn incrementally using boosting [32].

2.1.6. Simulation Study

A linear function is included to observe the comparison between the mixed-effects machine learning model and a linear mixed-effects model. This function is shown in Equation 1.

In addition, the nonlinear function 'Friedman3' is also included in the simulation study. The 'Friedman3' function was first introduced by [15] and is frequently used to compare nonparametric regression models (Equation 2).

$$y\,(x) = C^{*}\, x + 2 \qquad \text{(Equation 1)}$$

$$y\,(x) = C^{*} \arctan\left(\frac{x_1 x_2 - \frac{1}{x_1 x_3}}{x_0}\right) \qquad \text{(Equation 2)}$$

The constant C is chosen so that the variance of F(X) is equal to 1, as can be seen in previous equations, namely F(X) has the same power as the random effects.

$$y = F(X) + Zb + \varepsilon, b \sim N(0, \sum), \ \varepsilon_i \sim N(0, \sigma^2 I_n) \qquad \text{(Equation 3)}$$

Sample sizes of 5000, 10000 and 15000; group numbers were taken as 250, 500, 1000.

A single grouping variable was used. However, random effects can also be used, including hierarchically clustered, nested, crossed, and random coefficient effects.

In the simulation study, random effect variance was taken as 1, error variance as 1 and 4 separately.

2.1.6.1. Model parameters for GPBoost

Max_Depth: The value of the tree's branches extending downwards. In other words, it is the depth of the tree. It should be optimized to avoid over-learning. Too much branching causes over-learning and too little branching causes under-learning. The default parameter is -1.

Learning_rate: A value between 0-1 for scaling trees. A smaller value helps better predictive power. But, it will increase learning time and the possibility of over-learning. The default parameter is taken as 0.1.

Iterations: It shows the number of trees to be created. It is also used with the names "num_boost_round", "n_estimators", "num_trees", "num_iterations" in different algorithms . If it is taken too small, it can cause under-learning, and too much can cause over-learning. In addition, the increase in the number increases the training time. The default parameter is 32.

Early_stopping_rounds: It is the parameter used to prevent over-learning. After finding the most suitable step, number of trials should be specified. With the value of 100, the model performs 100 more iterations after the optimal, and the model stops learning even if the target parameters are not captured. For example, if the number of iterations is 2000 in the first parameter, if the optimal moment is reached in the 1000th iteration, the model will stop there. The default parameter is taken as 5.

Verbosity: In each iteration, the model's learning status, total time and remaining time are output. This output takes up too much space on the screen in case of multiple iterations and does not provide enough information to be worth it. The default parameter is set to 0.

feature_fraction: If less than 1.0, it randomly selects a subset of features (parameters) at each iteration (tree). For example, if taken as 0.8, the algorithm will select 80% of the parameters before training each tree. The default parameter is 1.0.

subsample: The subsample ratio of the training sample. It checks the samples given to the trees. For example, if set to 0.5 it means allowing the algorithm to randomly sample half of the training data before growing the trees, which will prevent over-learning. The default parameter is 1.0.

In order to make an objective comparison of the decision tree algorithm and boosting algorithms used, the basic parameters used in the algorithms were entered the same for each algorithm. Parameters with the same values are respectively; “n_estimators” are “learning_rate” and “max_depth”. A high max_depth value creates a more complex model and increases the likelihood of overfitting. In addition, the cross validation value operation was performed by determining the K-fold value as 4. The cross validation value process was used in all models in this thesis with the same parameter values. On the other hand, the 'n_jobs' parameter ensures that the calculation process is done using parallel processors, so that the processes are carried out faster. Taking the value of -1 for this parameter ensures that all CPUs are used [9].

2.1.6.1.1 Tuning parameters for GPBoost

Tuning parameters were chosen using 4-fold cross validation on the training data with RMSE as a criterion and ignoring random effects for GPBoost model predictions.

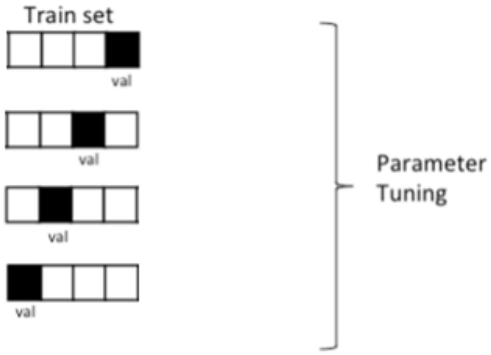


Figure 2: Parameter tuning for the training set using 4-fold cross validation [19]

2.1.6.2. Model parameters for MERF

Merf Python package (version 0.3) is used for the MERF algorithm. A maximum tree depth limit was not set and the number of trees was set to 300. These are the default values of the MERF package [23].

2.1.7. Error Measurement Parameters

2.1.7.1. Root Mean Square Error (RMSE)

The Root Mean Squared Error (RMSE) is the standard deviation of the residuals (estimation errors). The residuals are a measure of how far the regression line is from the data points. The RMSE is a measure of how far these residues have variated. In other words, it tells how dense the data is around the optimal line. Root mean square error is widely used in climatology, prediction and regression analysis to validate experimental results. RMSE is expressed by the following equation [1].

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad \text{(Equation 4)}$$

2.1.7.2 System Capacity

All remaining calculations are performed on a personal computer with an Intel Core i7 2.8 GHz quad-core processor and 8 GB of random access memory (RAM).

3. Results

Table 1: It shows the results for the grouped random effects model and the mean F = 'linear'
$\boldsymbol{\sigma_1^2}$: random effects variance $\boldsymbol{\sigma_\varepsilon^2}$: error variance

				Linear ME		**MERF**		**GPBoost**	
Number of Samples	**Number of Groups**	$\boldsymbol{\sigma_1^2}$	$\boldsymbol{\sigma_\varepsilon^2}$	**RMSE**	**time (s)**	**RMSE**	**time (s)**	**RMSE**	**time (s)**
5000	250	1	1	**1.241**	0.009	1.246	134.018	1.250	0.060
	500			**1.218**	0.016	1.224	183.194	1.234	0.068
	1000			**1.272**	0.016	1.283	277.043	1.284	0.110
5000	250	1	4	**2.136**	0.005	2.145	139.471	2.152	0.130
	500			**2.217**	0.002	2.134	185.186	2.240	0.113
	1000			2.235	0.002	**2.197**	285.063	2.258	0.147

Table 2: It shows the results for the grouped random effects model and the mean F = 'linear'

σ_1^2: random effects variance σ_ε^2: error variance

				Linear ME		MERF		GPBoost	
Number of Samples	**Number of Groups**	σ_1^2	σ_ε^2	**RMSE**	**time (s)**	**RMSE**	**time (s)**	**RMSE**	**time (s)**
10000	250	1	1	**1.240**	0.008	1.244	186.374	1.246	0.099
	500			**1.230**	0.015	1.235	214.008	1.236	0.098
	1000			**1.267**	0.012	1.270	316.856	1.273	0.128
10000	250	1	4	**2.146**	0.011	2.152	180.332	2.156	0.192
	500			**2.151**	0.016	2.160	216.040	2.162	0.187
	1000			2.252	0.016	**2.185**	336.068	2.268	0.194

Table 3: It shows the results for the grouped random effects model and the mean F = 'linear'
σ_1^2: random effects variance σ_ε^2: error variance

				Linear ME		MERF		GPBoost	
Number of Samples	**Number of Groups**	σ_1^2	σ_ε^2	**RMSE**	**time (s)**	**RMSE**	**time (s)**	**RMSE**	**time (s)**
15000	250	1	1	**1.232**	0.016	1.236	274.723	1.238	0.117
	500			**1.219**	0.036	1.222	246.556	1.224	0.154
	1000			**1.244**	0.016	1.245	316.872	1.246	0.166
15000	250	1	4	**2.118**	0.017	2.123	368.940	2.126	0.246
	500			**2.121**	0.016	2.126	238.088	2.130	0.270
	1000			**2.142**	0.016	2.144	337.542	2.147	0.314

Table 4: It shows the results for the grouped random effects model and the mean F = 'friedman3'
σ_1^2: random effects variance σ_ε^2: error variance

				Linear ME		MERF		GPBoost	
Number of Samples	**Number of Groups**	σ_1^2	σ_ε^2	**RMSE**	**time (s)**	**RMSE**	**time (s)**	**RMSE**	**time (s)**
5000	250	1	1	1.404	0.007	**1.256**	143.624	1.258	0.051
	500			1.400	0.004	**1.257**	184.849	1.257	0.062
	1000			1.441	0.062	**1.289**	282.554	1.290	0.092
5000	250	1	4	2.264	0.005	**2.176**	141.552	2.183	0.127
	500			2.264	0.012	**2.181**	183.743	2.188	0.120
	1000			2.331	0.002	**2.199**	282.643	2.261	0.154

Table 5: It shows the results for the grouped random effects model and the mean F = 'friedman3'
σ_1^2: random effects variance σ_ε^2: error variance

				Linear ME		MERF		GPBoost	
Number of Samples	**Number of Groups**	σ_1^2	σ_ε^2	**RMSE**	**time (s)**	**RMSE**	**time (s)**	**RMSE**	**time (s)**
10000	250	1	1	1.429	0.009	1.282	203.847	**1.275**	0.124
	500			1.389	0.014	1.236	241.786	**1.232**	0.116
	1000			1.401	0.016	1.252	329.883	**1.250**	0.146
10000	250	1	4	2.258	<0.001	**2.167**	204.314	2.170	0.188
	500			2.225	0.010	2.139	242.465	**2.138**	0.206
	1000			2.244	0.022	**2.159**	330.809	2.164	0.236

Table 6: It shows the results for the grouped random effects model and the mean F = 'friedman3'

$\boldsymbol{\sigma_1^2}$: random effects variance $\boldsymbol{\sigma_\varepsilon^2}$: error variance

				Linear ME		MERF		GPBoost	
Number of Samples	**Number of Groups**	σ_1^2	σ_ε^2	**RMSE**	**time (s)**	**RMSE**	**time (s)**	**RMSE**	**time (s)**
15000	250	1	1	1.390	0.016	1.253	314.325	**1.246**	0.131
	500			1.391	0.017	1.258	296.733	**1.249**	0.144
	1000			1.393	0.025	1.249	378.842	**1.243**	0.178
15000	250	1	4	2.223	0.031	**2.146**	310.047	2.147	0.249
	500			2.227	0.010	2.153	294.676	**2.151**	0.254
	1000			2.239	0.014	2.158	376.587	**2.158**	0.307

4. Conclusion

It is recommended to use generalized mixed-effects models to relax the linearity assumption in mixed-effects models. This may cause the model to be fitted incorrectly. For this reason, clustered or grouped random effects models for longitudinal data, non-parametric, machine learning-based approaches have been proposed. [38] merf model, [54] used the GPBoost model.

GPBoost machine learning algorithm is an effective method for both relax the linearity assumption in Gaussian process and mixed effects models, and relaxing the independence assumption for boosting.

In the case of linear models, coefficient estimates are less effective when fixed effects are used rather than random effects. Therefore, the mean function is likely to be estimated to be less efficient when using boosting for fixed effects [54].

p-values were also calculated from dependent t-tests comparing the GPBoost algorithm with other approaches. It has proven that the differences between the methods were significant statistically ($p<0.001$).

No matter how long it worked, the convergence of the MERF algorithm could not be observed most of the time. This may be related to the MERF algorithm not having a properly defined EM algorithm.

In the simulation step of the study, for a nonlinear function, GPBoost showed a better performance in terms of RMSE and time than MERF and LME methods. However, when a linear function is considered, LME gives a better result.

This method, in the field of medicine; the localization of gene and protein markers contributes to additional solution, in short, more informative and promising diagnostic decision-making about disease pathogenesis and etiology. In the field of robotics; it can perform effectively in motion planning, positioning, localization and mapping. In the field of sports; the use of past years for NBA free agency estimation can be continued [41].

Keywords: Machine Learning, Random Effects, Simulation

References

1. Anonymous. "Root Mean Square Error". Access Address: [https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/] Access Date: 10/10/2021.
2. A. M. Schmidt and P. Guttorp (2020). "Flexible Spatial Covariance Functions". Spatial Statistics, 100416.
3. A. Zeileis, T. Hothorn, K. Hornik (2008). "Model-based Recursive Partitioning, Journal of Computational and Graphical Statistics", 17 (2), 492-514.
4. Ashcroft, Michael. "Advanced Machine Learning: Gaussian Processes". Access Address: [https://futurelearn.com/info/courses/advanced-machine-learning/0/steps/49565]. Access Date: 07/04/2021.
5. Ateş, B (2020). "The Forecasting of Stress Concentration in Ship Buildings by Using Rough Mesh Structure and Machine Learning Method". Istanbul Technical University. Department of Shipbuilding and Ocean Engineering Shipbuilding and Ocean Engineering Program, Master`s Thesis.
6. Banerjee, S. et al. (2003). "Hierarchical Modeling and Analysis for Spatial Data". Chapman and Hall/CRC.
7. Bassetti, T. et al. F. (2018). "The Tree of Political Violence: a GMERT Analysis". Empirical Economics, 54(2), 839-850.
8. Cochrane, C. et al. (2021). "An Ensemble Mixed Effects Model of Sleep Loss and Performance. Journal of Theoretical Biology". 509, 110497.
9. Coşkun, K. (2020). "Comparison of the Decision Tree Based Boosting Algorithms in Classification of Network Attacks". Muğla Sıtkı Koçman University. Institute of Sciences, Department of Information Technologies Engineering, Master`s Thesis.
10. Cressie, N. and Wikle, C. K. (2015). "Statistics for Spatio-Temporal Data". John Wiley & Sons.
11. De'Ath, G. (2002). "Multivariate Regression Trees: A New Technique for Modeling Species–Environment Relationships." Ecology, 83(4), 1105-1117.
12. DeBruine, L. M. and Barr, D. J. (2020). "Understanding Mixed Effects Models Through Data Simulation". Advances in Methods and Practices in Psychological Science.
13. Fokkema, M. et al. (2018). "Detecting Treatment-Subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees". Behavior Tesearch Methods, 50(5), 2016-2034.
14. Fontana, L. et al. (2021). "Performing Learning Analytics via Generalised Mixed-Effects Trees Data", 6(7), 74.
15. Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines". The Annals of Statistics, 1-67.
16. Fu, W. and Simonoff, J. S. (2015), "Unbiased Regression Trees for Longitudinal and Clustered Data". Computational Statistics & Data Analysis, 88, 53-74.
17. G.-A. Fuglstad et al. "Does Non-Stationary Spatial Data Always Require Non-Stationary Random fields?" (2015). Spatial Statistics, 14:505-531.
18. Galeshchuk, S. and Qiu, J. (2019). "Identification of Opinion Makers on Twitter. in International Conference on Data Science and Social Research" (pp. 187-198). Springer, Cham.
19. Gamze E. K. (2020). "Prediction of Druggability Properties of Chemicals Using Machine Learning Techniques". Boğaziçi University. Institute of Sciences, Department of Computer Engineering, Master`s Thesis.
20. Gneiting, T. et al. (2007). "Probabilistic Forecasts, Calibration and Sharpness". Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243-268.
21. H. Zhang (1998). "Classiffication Trees for Multiple Binary Responses". Journal of the American Statistical Association, 93 (441), 180-193.
22. Hajjem, A. et al. (2011). "Mixed Effects Regression Trees for Clustered Data". Statistics & Probability Letters, 81(4), 451-459.
23. Hajjem, A. et al. (2014). "Mixed-Effects Random Forest for Clustered Data". Journal of Statistical Computation and Simulation, 84(6), 1313-1328.

24. Hajjem, A. et al. (2017). "Generalized Mixed Effects Regression Trees". Statistics & Probability Letters, 126, 114-118.
25. Hothorn, T. et al. (2010). "Model-based Boosting 2.0. Journal of Machine Learning Research". 11, 2109-2113.
26. J. H. Friedman (2001). "Greedy Function Approximation: a Gradient Boosting Machine". Annals of Statistics, 1189-1232.
27. Jones, D. R. et al. (1998). "Efficient Global Optimization of Expensive Black-box Functions". Journal of Global Optimization, 13(4), 455-492.
28. Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian Calibration of Computer models". Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3), 425-464.
29. Knagg, O. (2019). "An Intuitive Guide to Gaussian Processes". Access Address: [https://towardsdatascience.com/an-intuitive-guide-to-gaussian-processes-ec2f0b45c71d]. Access Date: 07/12/2020.
30. Kükner, C.U. (2020). "A Comparative Analysis of LSTM and XGBOOST Methods for Day Ahead Electricity Price Forecasting". Istanbul Technical University. Energy Science and Technology Institute, Department of Energy Science and Technology, Master`s Thesis.
31. L. Breiman (2001). "Random Forests", Machine Learning, 45 (1), 5-32.
32. L. Breiman, J. Friedman et al. (1984). "Classiffication and Regression Trees", CRC Press.
33. L. Breiman (1998). "Arcing Classiffiers". Annals of Statistics, pages 801-824.
34. Laird, N. M. and Ware, J. H. (1982). "Random-effects Models for Longitudinal data". Biometrics, 963-974.
35. Levy, J. et al. (2021). "Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction Using Spatially Localized Immuno-Oncology Markers". bioRxiv.
36. M. R. Segal (1992). "Tree-structured Methods for Longitudinal Data". Journal of the American Statistical Association, 87 (418) 407-418.
37. Matheron, G. (1963). "Principles of Geostatistics". Economic Geology, 58(8), 1246-1266.
38. merf python package (v0.3; Manifoldai, 2019).
39. Mukadam, M. et al. (2016). "Gaussian Process Motion Planning". In 2016 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9-15). IEEE.
40. N. Cressie and C. K. Wikle. (2015) "Statistics for Spatio-temporal Data". John Wiley & Sons.
41. Natarajan, S. S. (2019). "A Mixed Effects Modeling Approach to Predicting NBA Free Agency". Access Address: [https://hockeygraphsdotcom.files.wordpress.com/2019/09/mixedfa_model_deck-senthil-v2.pdf]. Access Date: 07/09/2020.
42. Ngufor, C. et al. (2019). "Mixed Effect Machine Learning: a Framework for Predicting Longitudinal Change in Hemoglobin A1c". Journal of Biomedical Informatics, 89, 56-67.
43. Pinheiro J, Bates D. (2000). "Mixed-Effects Models in S and S-PLUS". Springer-Verlag, New York.
44. Python, I. (2019). Python.
45. R Core Team (2013). R: A Language and Environment for Statistical Computing, [http://www.r-project.org/].
46. R. H. Shumway and D. S. Stoffer (2017). "Time Series Analysis and its Applications: with R Examples". Springer.
47. R. J. Sela and J. S. Simonoff (2012). "RE-EM trees: a Data Mining Approach for Longitudinal and Clustered Data". Machine Learning, 86 (2) 169-207.
48. S. Banerjee, B. P. Carlin and A. E. Gelfand (2014). "Hierarchical Modeling and Analysis for Spatial Data". Chapman and Hall/CRC.
49. Schapire, R. E. (1999). "A Brief Introduction to Boosting". In Ijcai (Vol. 99, pp. 1401-1406).
50. Segal, M. R. (1992). "Tree-structured Methods for Longitudinal Data". Journal of the American Statistical Association, 87(418), 407-418.
51. Sela, R. J. and Simonoff, J. S. (2009). "RE-EM trees: a New Data Mining Approach for Longitudinal Data". NYU Stern Working Paper SOR-2009-03.
52. Sela, R. J. and Simonoff, J. S. (2011). REEMtree: Regression Trees with Random Effects. R Package version 0.90, 3, 741-749.
53. Shumway, R. H. et al. (2000). "Time Series Analysis and its Applications (Vol. 3)". New York: Springer.
54. Sigrist, F. (2020). "Gaussian Process Boosting". arXiv preprint arXiv:2004.02653.

55. Snoek, J. et al. (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". Advances in Neural Information Processing Systems, 25.
56. Speiser, J. L. et al. (2020). "BiMM Tree: A Decision Tree Method for Modeling Clustered and Longitudinal Binary Outcomes". Communications in Statistics-Simulation and Computation, 49(4), 1004-1023.
57. Torsten Hothorn, et al. (2010). "mboost: Model-Based Boosting", [http://r-forge.r-project.org/projects/mboost.] R package version 2.1-0.
58. Verbeke G. and Molenberghs G. (2000). "Linear Mixed Models for Longitudinal Data". New York John Wiley Sons.
59. W. W. Stroup (2012), "Generalized Linear Mixed Models: Modern Concepts, Methods and Applications". CRC Press.
60. Williams, C. K. and Rasmussen, C. E. (2006). "Gaussian Processes for Machine Learning" (Vol. 2, No.3, p. 4). Cambridge, MA: MIT Press.
61. Wu, H. and Zhang, J. T. (2006). "Nonparametric Regression Methods for Longitidunal Data Analysis: Mixed-effects Modeling Approaches". Wiley, New York.
62. Yangın, G. (2019), "Application of XGboost and Decision Tree Based Algorithms on Diabetes Data". Mimar Sinan Fine Art University. Institute of Sciences, Master`s Thesis.
63. Yılmaz, H. (2014). "Studying the Missing Data Problem in Random Forests Method and an Application in Health Field". Eskişehir Osmangazi University. Institute of Health Sciences, Department of Biostatistics, Master`s Thesis.
64. Zeileis, A. et al. (2008). "Model-based Recursive Partitioning". Journal of Computational and Graphical Statistics, 17 (2), 492–514.

S10

Comparison of Some Classification Methods in Ordinal Data Structure by Simulation

Ali Vasfi AĞLARCI[1], Cengiz BAL[2]

1.Eskişehir Osmangazi University, Faculty of Medicine, Department of Biostatistics, Eskişehir (PhD Student)
1.Bartın University, Quality Coordinator, Bartın (Lecturer)
2.Eskişehir Osmangazi University, Faculty of Medicine, Department of Biostatistics, Eskişehir (Assoc. Prof.)

Aim: This study aims to examine the classification performances of classification methods (Support Vector Machines (DVM), K-Nearest Neighbor (KNN), Ordinal Logistic Regression (OLR), Artificial Neural Networks (ANN), Random Forest (RF), Classification and Regression Trees (CART)) when the response variable is in ordinal structure.

Method: Data sets in this research were derived using the R package program, 1000 repetitive simulation studies were carried out with each method under various scenarios, and the classification results of the methods were examined. With the simulation study, 3 different correlation structures (low, medium, high), 3 different ordinal response variable categories (3, 4 and 5 categories), 5 different sample sizes (100, 250, 500, 1000, 2500) and response variable Classification performances of the above-mentioned methods were evaluated by deriving 10 variables (1 ordinal, 3 binary, 6 continuous) for 90 different situations with or without equal distribution. 75% of the data derived in the classification process with these methods was used as training data and 25% as test data. While evaluating the performances of the methods, interpretations were made considering the kappa value.

Findings and Conclusion: As a result of the evaluations, it was seen that the classification performance of the methods increased in all scenarios as the level of correlation increased. The change in sample size affected the classification performance of the methods differently. In classification with SVM, KNN, RF methods, it was observed that the classification performance increased or tended to increase as the sample size increased in all scenarios. In the classification with the OLR method, in cases where the response variable has 4 and 5 categories in the low correlation structure and is not evenly distributed, it has been observed that the increase in the sample size decreases the classification performance of the method and increases the performance in all other scenarios. In the ANN method, on the other hand, it was observed that the classification performance decreased as the sample size increased. In the applications made with the CART algorithm, it was observed that the classification performance of the method increased up to 500 sample size, and the performance began to decrease with the gradual increase towards 1000 and 2500 sample sizes. Considering 90 different scenarios as a result of the simulation study, it was seen that the SVM method classifies better than other methods when the response variable is ordinal. Meanwhile, the CART showed lower classification performance when compared to other algorithms.

Keywords: Classification, Ordinal Data, Machine Learning, Ordinal Logistic Regression

References

1) Aydın Haklı, D., Başol, M., Öztürk, E., & Karabulut, E. (2017). Effectiveness of Three Factors on Classification Accuracy. *10th International Statictics Congress.*

2) Dag, O., Kasikci, M., Karabulut, E., & Alpar, R. (2019). Diverse classifiers ensemble based on GMDH type neural network algorithm for binary classification. *Erişim tarihi:20.04.2020, https://www.tandfonline.com/doi/full/10.1080/03610918.2019.1697451*.

3) Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. . *Journal of biomedical informatics, 35(5-6),* 352-359.

4) Dirican, E. (2019). Makine Öğrenimi Yöntemlerini Kullanarak Evre III İnvaziv Duktal Karsinomlu Hasta Verilerinin Sınıflandırılması . *Doktora Tezi, Dicle Üniversitesi Sağlık Bilimleri Enstitüsü, Diyarbakır*.

5) Elasan, S. (2019). Veri Madenciliğinde Farklı Karar Ağaçları Ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları Ve Doğum Verisinde Bir Uygulama . *Doktora Tezi, Van Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü, Van* .

6) Fabacher, T., Godet, J., Klein, D., M, V., & Jegu, J. (2020). Machine learning application for incident prostate adenocarcinomas automatic registration in a French regional regional cancer registry. *International Journal of Medical Informatics, Volume 139*.

7) Heo, J., Park, S., Kang, S., Oh, C., Bang, J., & Kim, T. (2020). Prediction of Intracranial Aneurysm Risk using Machine Learning. *Nature Research, Scientific Reports volume 10, Article number: 6921*.

8) Huang, C., Yang, Y., Yang, D., & Chen, Y. (2009). Frog classification using machine learning techniques,. *Expert Systems with Applications*.

9) Köktürk, F. (2012). K-En Yakın Komşuluk, Yapay Sinir Ağları Ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması . *Doktora Tezi, Bülent Ecevit Üniversitesi Sağlık Bilimleri Enstitüsü, Zonguldak* .

10) Lin, G., Nagamine, M., Yang, S., Tai, Y., Lin, C., & Sato, H. (2020). Machine Learning Based Suicide Prediction for Military Personnel. *Journal of Biomedical and Health Informatics*.

11) Mbiki, S., McClendon, J., & Gilmore, J. (2020). Classifying changes in LN-18 glial cell morphology: a supervised machine learning approach to analyzing cell microscopy data via FIJI and WEKA. *Medical & Biological Engineering & Computing*.

12) Moon, S., Song, P., Sharma, V., Lyons, K., Pahwa, R., Akinwuntan, P., & Devos, P. (2020). Classification of Parkinson's disease and essential tremor based on gait and balance characteristics from wearable motion sensors: A data-driven approach. *The Preprint Server For Healty Sciences*.

13) O'Connell, A. (2006). *Logistic Regression Models for Ordinal Response Variables.* . London: Thousand Oaks, Sage Publications .

14) Wang, Y., Yu, C., & Chan, H. (2012). Predicting Construction Cost and Schedule Success Using Artificial Neural Networks Ensemle and Support Vector Machines Classification Models. *International Journal of Project Management*, 470-478.

15) Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient knn classification with different numbers of nearest neighbors . *IEEE transactions on neural networks and learning systems, 29(5),*, 1774-1785.

S11

Predictive Performance of Expectation Maximization, K-Nearest Neighbors and Random Forests Imputation Methods for Propensity Score Matching in Missing Data Problem

Buğra VAROL[1], İmran KURT ÖMÜRLÜ[2], Mevlüt TÜRE[2]

[1] *Adnan Menderes University, Institute of Health Sciences, Division of Biostatistics, Aydın*
[2] *Adnan Menderes University, Faculty of Medicine, Division of Biostatistics, Aydın*
**Presenter e-mail: bugravarol87@gmail.com*

Introduction: Propensity score (PS) is an effective method to control for covariates in observational studies. A challenge in PS analyses are missing values in covariates. The aim of this study is to investigate how different three imputation methods of handling missing values of covariates in a PS analysis can affect average treatment on the treated (ATT) estimates.

Method: In this study, missing data imputation methods were evaluated using different data sets, whose covariates were low, medium, and high (r=0.10, 0.50, 0.85) correlated with each other, for n=200 units and 1000 times running simulation. Missing data structures were created according to the missing at random (MAR) mechanism and different missing rates. Different datasets were obtained after having imputed the missing values separately by three imputation methods including expectation maximization (EM), k-nearest neighbors (KNN) and random forests (RF). Then the PS nearest neighbor matching was implemented and ATT scores obtained using the imputed data sets.

Results: ATT scores that obtained from EM, KNN and RF methods ranged from 3.284 to 3.786 for low correlated data sets, 2.941 to 3.250 for medium correlated data sets and, 2.563 to 2.702 for high correlated data sets. The RF method was found as the method having the lowest bias value and its prediction performance showed statistical superiority compared to the other two methods for low, medium and high correlated data sets. Additionally, almost all imputation methods tended to produce lower bias values when less data was missing. As the rate of missing data increased, the bias also increased.

Conclusion: Ignoring missing values on covariates for PS analyses causes information loss significantly and this information loss becomes greater as the rate of missing data increases. PS analyses might be biased if missing data on covariates are ignored also. To prevent this information loss and bias, PS analyses should be performed after solving the problem of missing data on covariates by RF methods, which showed statistical superiority compared to other two methods in this study.

Key words: Missing Data, Simulation, Propensity Score, Random Forests, Expectation Maximization, K-Nearest Neighbors .

References

1. Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological methods*, *21*(3), 427.

2. Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European journal of epidemiology*, *34*(1), 23-36.

3. D'Agostino Jr, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, *95*(451), 749-759.

4. Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913-933.

5. Jochen, H., Max, H., Tamara, B., & Wilfried, L. (2013). Multiple imputation of missing data: a simulation study on a binary response. *Open Journal of Statistics*, *2013*.

6. Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC bioinformatics*, *20*(1), 1-11.

7. Le, T. D., Beuran, R., & Tan, Y. (2018, November). Comparison of the most influential missing data imputation algorithms for healthcare. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 247-251). IEEE.

8. Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

9. Misztal, M. (2012). Imputation of missing data using R package.

10. Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, *179*(6), 764-774.

S12

Comparison of Different Machine Learning Methods in the Diagnosis of Chronic Renal Failure (CRF)

Tolga DEMİREL[1]. Associate Professor Doctor Leman TOMAK[1]

[1]*Ondokuz Mayıs University. Faculty of Medicine. Department of Biostatistics and Medical Informatics. SAMSUN*

Abstract

Aim: Chronic Renal Failure (CRF) is a health problem that affects the social and psychological states of individuals and machine learning methods have an important place in its diagnosis. The aim of this study is to evaluate prediction models with different machine learning algorithms in the diagnosis of CRF and to compare the performances of these models.

Methods: 25 different variables related to a total of 400 people were used in the data set related to the diagnosis of CRF taken from the University of California Machine Learning Repository (UCI-ML). Of all, 250 people have CRF whereas 150 people do not have this disease. WEKA software was used in order to analyze the data set and OneR, Naive Bayes, decision trees, logistic regression, support vector machines and nearest k-neighbor algorithms were modelled. The success rates of these models were compared using the accuracy rate criterion.

Results: In this study, the algorithm with the highest performance was Decision Trees with an accuracy rate of 99% whereas the algorithm with the lowest accuracy rate was the OneR Algorithm with a rate of 92%. Accuracy rates of other algorithms was determined as 97.75% for support vector machines, 95.75% for nearest k-neighbor, 95.75% for logistic regression and 95% for Naive Bayes.

Conclusions: In this study, when the accuracy rate criteria for the algorithms used in the diagnosis of CRF are compared, the most successful algorithm model is decision trees with an accuracy rate of 99% and the accuracy rate for other machine learning algorithms is over 90%. In line with these results, it can be said that machine learning methods are very successful predictive models in the diagnosis of CRF.

Keywords: Machine learning, machine learning algorithms, chronic renal failure, accuracy rate.

Introduction

Machine learning refers to computer science techniques used to give machines the ability to learn without the need for explicit programming. To put it more clearly, it is all of the teaching techniques that rely on statistical methods to make computers a structure that can think and analyze like humans [1]. In this way, people become able to analyze large and complex data that cannot be analyzed.

Machine learning is used in many sectors and fields, especially health, informatics, finance, industry, production, agriculture and education. The question of how systems can be automatically adapted or optimized to their environment is an important research topic for which human beings seek its answer. Machine learning is an important data mining method that is used to answer this question and has become widespread recently. Thanks to machine learning, significant developments have been achieved in different fields. For example, fraud detection in the field of cyber security, identification of dangerous and unwanted e-mails detection of network security threats and presentation of product advertisements according to preferences in the field of marketing could be achieved by using different machine learning methods [2].

There are different algorithms developed for accurate classification predictions in machine learning. Among these algorithms, OneR, linear regression, logistic regression, Naive Bayes, support vector machines, k-nearest neighbor, decision trees, random forest, random tree algorithms are the most widely used. The common purpose of these algorithms is to create the statistical model that helps to predict most accurately the variable that wanted to be classified by using the information from other variables, and improve by learning the model created with the data constantly added to the system [3].

Chronic renal failure (CRF) disease can be defined as a chronic and progressive deterioration state in adjusting the fluid-solute balance of the kidney and metabolic-endocrine functions as a result of decreased glomerular filtration value. Kidney size have reduced in most of the CRF patients. The clinical signs and symptoms of the patients are closely related to the underlying pathology, the degree of renal failure and the rate of development [4, 5]. In addition to its medical aspect, CRF disease also affects the social, economic and psychological conditions of the patients.

In the 2019 Turkish Kidney Registration System Report of the Turkish Society of Nephrology, it was stated that there were 61341 hemodialysis, 3292 peritoneal dialysis, and approximately 19150 organ transplants including pediatric cases in total 83783 kidney failure cases in Turkey [6]. Many techniques and models have been developed to diagnose this disease early. There are different machine learning methods among them. Machine learning methods are effectively used in the early diagnosis of various diseases and in the treatment process in the field of health, as in many areas [7].

The aim of this study is to create prediction models with different machine learning algorithms in the diagnosis of CRF and to evaluate the performances of these models by comparing them.

Method

Study Data

The data used in this study was taken from the University of California Machine Learning Repository (UCI-ML). In the data set of chronic renal failure disease, 25 different variables belonging to 400 people were used (Table 1). While 250 of the 400 people in the data set have CRF, 150 of them do not have this disease [8]. WEKA software developed by the University of Waikato was used in this study.

Table 1. Data characteristics used in the study

Sequence No	Variable	Definition	Measurement and Range Value	Numeric/Categorical
1	Age	Age year	Age of patient	Numeric
2	bp	Blood plesure	mm/Hg	Numeric
3	sg	Specific gravity	1.005 - 1.010 - 1.015 - 1.020 - 1.025	Categorical
4	al	Albumin	0 - 1 - 2 - 3 - 4 - 5	Categorical
5	su	Sugar	0 - 1 - 2 - 3 - 4 - 5	Categorical
6	rbc	Red blood cells	normal/abnormal	Categorical
7	pc	Pus cell	normal/abnormal	Categorical
8	pcc	Pus cell clumbs	available/not available	Categorical
9	ba	Bacteria	available/not available	Categorical
10	bgr	Blood glucose random	mgs/dl	Numeric
11	bu	Blood urea	mgs/dl	Numeric
12	sc	Serum creatinine	mgs/dl	Numeric
13	sod	Sodium	mEq/L	Numeric
14	pot	Potassium	mEq/L	Numeric
15	hemo	Hemoglobin	gms	Numeric
16	pcv	Packed cell volume	numeric value	Numeric

17	wc	White blood cell	cells/cumm	Numeric
18	rc	Red blood cell count	millions/cmm	Numeric
19	htn	Hypertension	yes/no	Categorical
20	dm	Diabetes mellitus	yes/no	Categorical
21	cad	Coronary artery disease	yes/no	Categorical
22	appet	Appetite	good/bad	Categorical
23	pe	Pedal edema	yes/no	Categorical
24	ane	anemi	yes/no	Categorical
25	class	Class	CRF / NCRF	Categorical

Methods in Classification

As with other prediction models, classifications made with statistical models created with machine learning algorithms aim to find the closest prediction to the truth. In this study, OneR, Naive Bayes, decision trees, logistic regression, support vector machines and nearest k-neighbor algorithms are used [9].

OneR Algorithm: In this algorithm, it is existed variable that gives the best result from the variables and a classification is made using only this variable. In this method, an error matrix is created for each variable and the accuracy rates in these matrices are calculated and the classification is made by using the variable that gives the best result [10].

Naive Bayes Algorithm: In this algorithm, the probability of the values received in all variables for each patient is calculated and the result in the target variable is estimated. Thus. the percentage of accuracy is calculated with the estimates found for all patients [5, 10].

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (1)$$

Decision Trees (J48) Algorithm: This algorithm is a frequently used data mining approach because it is advantageous for decision makers in terms of easy interpretation and comprehensibility. In the general approach, in decision trees the entropies belonging to the variables are calculated with Shannon's Theorem. Decision tree with maximum entropy value calculated from variables composes top branch of the decision tree [1, 5].

Entropy calculation formula for each variable;

$$\sum_{i=1}^{n} P_i . Log_2(P_i) \quad (2)$$

P_i: The relevant variable i. is the probability value of the category.

By making these calculations for each variable, the classification is shaped down as a tree branch.

Logistic Regression: It is an algorithm model that can be used when the variable to be predicted is categorical. In linear regression, the parameters of the independent variables are estimated by the least squares method while in logistic regression, the parameters are calculated by the maximum likelihood probability method. The aim of the maximum likelihood probability method is to select the best parameters that maximize the probability of occurrence of the data set from the infinite parameter pool. Since there is more than one independent variable in the data set used in this study, multiple logistic regression, the formula of which is given below was used [5, 9, 11].

$$Logit[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (3)$$

Support Vector Machines (SVM): This algorithm, similar to the logistic regression algorithm aims to create the best line or shape that separates the categories of the target variable. It is known that in many applications, support

vector machines outperform traditional machine learning methods and are shown as a powerful tool for solving classification problems [5, 9].

Nearest K-Neighbor (K-NN): In this algorithm, there are k people closest to the person estimated during classification, and the relevant person is included in that class, just as the majority of these people are classified. If k is taken as 1, the patient to be estimated is included in the class of the closest patient. While calculating the proximity distance between neighbors, usually the euclidean measure is used [3, 5].

Performance Criteria

The indicators used to evaluate the performance of machine learning algorithms are called performance criteria. Accuracy, sensitivity, specificity and susceptibility are the most commonly used performance criteria [1, 5, 11]. In this study, accuracy ratio was used as performance criteria. The accuracy rate is calculated with divide the number of all categories that the algorithm used correctly predicts in the target variable by the total number of people. The accuracy rate is obtained using the confision matrix shown in Table 2.

Table 2. Confision matrix

Confision Matrix		Accuracy Rate
T(CRF)	T(NCRF)	**[T(CRF)+T(NCRF)] / N**
F(CRF)	F(NCRF)	

Given in the confision matrix;

N : Total number of people
CRF : People with chronic renal failure
NCRF : People without chronic renal failure
T(CRF) : People with correctly predicted and chronic renal failure
F(CRF) : People with wrong predicted and chronic renal failure
T(NCRF) : People who are correctly predicted and do not have chronic renal failure
F(NCRF) : People who are wrong predicted and do not have chronic renal failure disease

Results

Accuracy rates with respect to algorithms are given in Table 3. The results of the study revealed that accuracy rates were computed as 92% for OneR, 95% for Naive Bayes, 99% for decision trees, 95.75% for logistic regression, 97.75% for support vector machines and 95.75 for nearest k-neighbor.

Table 3. Comparison of machine learning algorithms

Sequence No	Algorithms	T(CRF)	F(CRF)	T(NCRF)	F(NCRF)	Accuracy Rate [T(CRF)+T(NCRF)]/N
1	OneR	229	11	139	21	92.00%
2	Naive Bayes	230	0	150	20	95.00%
3	Decision Trees (J48)	249	3	147	1	99.00%
4	Logistic Regression	238	5	145	12	95.75%
5	Support Vector Machines (SVM)	241	0	150	9	97.75%
6	Nearest K-Neighbor	233	0	150	17	95.75%

Graphical representation of accuracy rates is presented in Figure 1.

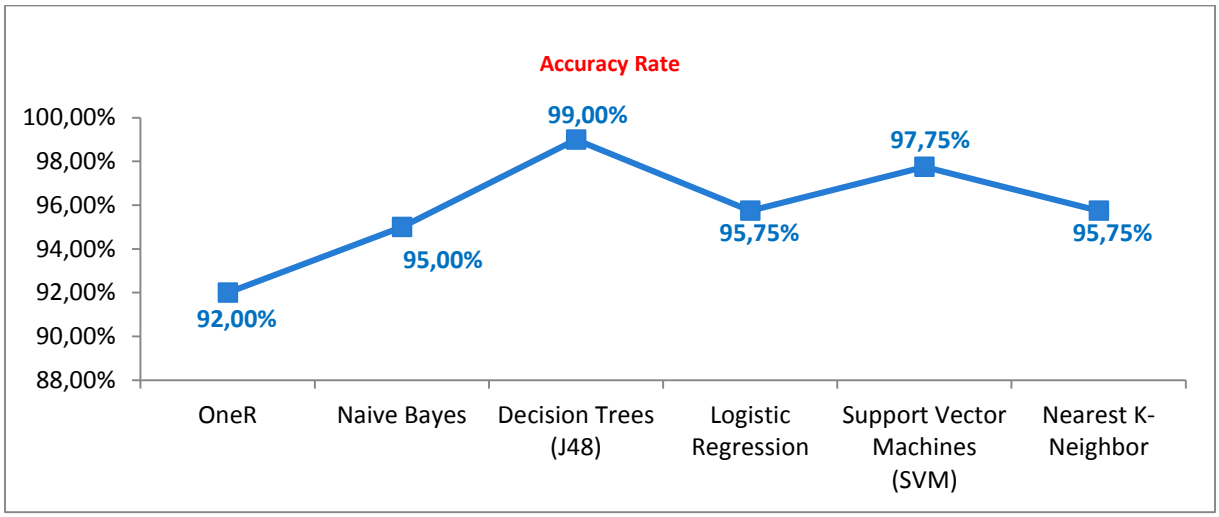


Figure 1. Accuracy rates of algorithms

Discussion and Conclusion

In this study, machine learning methods were compared in the diagnosis of CRF and, the accuracy rate was used as a performance criterion. The results of the study highlighted that the highest performance was achieved by the decision tree algorithm with an accuracy rate of 99%. This algorithm is respectively followed by support vector machines with an accuracy of 97.75%, logistic regression and nearest k-neighbor algorithms with an accuracy of 95.75%, Naive Bayes with an accuracy of 95%, and finally a OneR algorithm with an accuracy of 92%.

Indicators and results of some academic studies similar to this study are shown in Table 4 [12, 13, 14, 15, 16].

Table 4. Similar academic studies and their results [12, 13, 14, 15, 16].

"Prediction and Diagnosis of Heart Disease by Data Mining Techniques" 2015	
Decision Trees (J48)	**83.73%**
Nearest K-Neighbor	82.78%
Support Vector Machines (SVM)	82.78%
Naive Bayes	81.82%

"Comparison of Machine Learning Algorithms in WEKA" 2017	
Naive Bayes	71.70%
Decision Trees (J48)	**75.50%**
Multi-Layer Perceptron (MLP)	64.70%

"Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients" 2015	
Naive Bayes	74.40%
Decision Trees (J48)	**81.80%**
Multi-Layer Perceptron (MLP)	82.30%

"Comparative Analysis of Classification Algorithms Using Weka" 2018	
Naive Bayes	95.00%
Decision Trees (J48)	**99.00%**
Multi-Layer Perceptron (MLP)	99.75%
Random Tree	95.50%

"Comparison of decision trees used in data mining" 2019	
Decision Trees (J48)	**67.79%**
Random Tree	62.92%
REP Tree	67.28%
Hoefding Tree	69.65%
Logistic Regression	70.33%

As seen in the studies given in Table 4, when the machine learning algorithms working with similar data sets are examined, the decision tree algorithm is generally among the algorithms that produce the best predictions and have a high accuracy rate. It was not found very large deviations in terms of accuracy rate between the machine learning methods used both in our study and the similar academic studies we have given as examples. In other words, the accuracy rates of the algorithms used generally give results close to each other.

In addition, among the different machine learning methods shown in Table 4, decision trees, Naive Bayes, multilayer perceptron algorithms were used more frequently. The differences between the accuracy rates of all algorithms used are about 10% at most. Therefore, it can be stated that the different algorithms can create models which are close to each other with an accuracy rate.

All six machine learning algorithms used in this study have an accuracy rate over 90%. In line with this result, it is possible to say that the models established by the machine learning methods are successful predictors in the diagnosis of chronic kidney failure disease.

As a result of these and similar studies, it can be seen that machine learning algorithms can be developed further and different machine learning algorithms can be created that can produce predictions much closer to reality. As a result of the academic resources created on this subject and the developing technological advances, it will be possible to take steps that will keep the costs at the lowest level, improve the ability to make quick decisions, and minimize the risk factor in many areas, including the health field.

References

1. B. Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), Jun 2020, pp.381-386.

2. T.M. Mitchell, "The Discipline of Machine Learning," Machine Learning Department School of Computer Science Carnegie Mellon University, July 2006.

3. Giuseppe Bonaccorso, "Machine Learning Algorithms," Packt Publishing Ltd., 2017.

4. M.H. Tanrıverdi, A. Karadağ, E.Ş. Hatipoğlu, "Chronic Renal Failure," Konuralp Medical Journal, 2010, pp.27-32.

5. B. Khan, R. Naseem, F. Muhammed, G. Abbas, S. Kim, "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy," IEEE ACCESS, March 2020, pp. 55012-55022.

6. G. Süleymanlar, K. Ateş, N. Seyahi, "Nephrology, Dialysis and Transplantation in Turkey 2019" Turkish Society of Nephrology Publications, 2020.

7. J.A. Roth, G. Radevski, C. Marzolini, A. Rauch, H.F. Günthard, R.D. Kauyos, C.A. Fux, A.U. Scherrer, A. Calmy, M. Cavassini, C.R. Kahlert, E. Bernasconi, J. Bogojeska, B. Battegay, "Cohort-Derived Machine Learning Models for Individual Prediction of Chronic Kidney Disease in People Living With Human Immunodeficiency Virus: A Prospective Multicenter Cohort Study," The Journal of Infectious Diseases, Oct 2020

8. URL-1, https://archive.ics.uci.edu/ml/machine-learning-databases/00336/, Sept 2021

9. R.S.G. Sealfon, L.H. Mariani, M. Kretzler, O.G. Troyanskaya, "Machine learning, the kidney, and genotype–

phenotype analysis," Kidney International, 2020, pp. 1141-1149.

10. V.S. Parsania, N.N. Jani, N.H. Bhalodiya, "Applying Naive Bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis," International Journal of Darshan Institute on Engineering Research and Emerging Technology, Vol. 3, No. 1, 2014, pp. 60-64.

11. A. Agresti, "An Introduction to Categorical Data Analysis," Aug 2006.

12. B.Bahrami, M.H.Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology (JMEST), Feb 2015, pp.164-168.

13. C.Almeida, M.Gonsalves, R.Manimozhi, "Comparison of Machine Learning Algorithms in WEKA," International Journal of Latest Trends in Engineering and Technology Special Issue SACAIM 2017, pp. 225-229.

14. K.J.Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients," researchgate.net, Apr. 2015.

15. S.Saini, A. Dhankkar, K.Solanki, "Comparative Analysis of Classification Algorithms Using Weka," IOSR Journal of Engineering (IOSRJEN), Oct. 2018, pp.29-40.

16. G. Aksu, N. Doğan, "Comparison of decision trees used in data mining," Pegem Journal of Education and Training, 2019, pp.1183-1208.

S13

Survival Prediction with Supervised Principal Components Analysis, Penalized Cox Models and Extreme Learning Machine based Penalized Cox Models on Simulated High Dimensional Data

Fulden CANTAŞ[1], İmran KURT ÖMÜRLÜ[1], Mevlüt TÜRE[1]

[1]*Adnan Menderes University, Faculty of Medicine, Division of Biostatistics, Aydın*

Introduction: The goal of this study is to compare classification performances of supervised principal components analysis (SuperPC), L2-penalized Cox model (Cox-L2) model and extreme learning machine-based L2-penalized Cox model (ELMCox) on long-short term survival prediction on high dimensional survival data.

Method: SuperPC, Cox-L2 and ELMCox methods are used in order to analyze high dimensional survival data. For this purpose, high dimensional breast cancer gene expression survival data are simulated by varying censoring rates in R programming language. At the end of the 1000 times-repetitive simulation process, accuracy rate, area under receiver operating characteristic curve (AUC), F1 score and Matthews correlation coefficient (MCC) are calculated as a performance metric.

Results: Median value of accuracy rates are found to range between 0.783 and 0.800 for SuperPC; 0.800 and 0.833 for Cox-L2; 0.783 and 0.817 for ELMCox method. Median AUC values of SuperPC, Cox-L2 and ELMCox methods are found in between 0.816 – 0.845; 0.843 – 0.866 and 0.811 – 0.863; respectively. The median value of F1 score is found to be changing from 0.769 to 0.800 for SuperPC method; from 0.800 to 0.824 for Cox-L2 method; from 0.776 to 0.820 for ELMCox method. The median value of MCC changes through 0.569 - 0.606 for SuperPC method; 0.546 - 0.645 for Cox-L2 method and 0.594 - 0.635 for ELMCox method.

Conclusion: As a consequence; Cox-L2 and ELMCox which give opportunity to analyze high-dimensional survival data without necessity to dimension reduction show competing classification performance to frequently used method SuperPC that analyzes high dimensional survival data after dimension reduction.

Keywords: Extreme Learning Machine, Supervised Principal Component Analysis, Penalized Cox Regression, Survival, Simulation, Classification

S14

Identifying the Factors Affecting the Intensive Care Unit Admission of Covid-19 Patients by Using Multivariate Statistical Methods

Neslihan Gökmen İnan[1], Arzu Baygül Eden[2]

[1]*Istanbul Technical University, Basic Sciences Department, Istanbul*
[2]*Koc University, Biostatistics Department, Istanbul*

Abstract

Aim: This study concerns with identifying intensive care unit (ICU) admission of COVID-19 cases on the first symptoms and blood-test results of admitted to Emergency Department of Koç University Hospital (Istanbul, Turkey). The aim of the study is to figure out important factors affecting ICU admission.

Method: The logistic regression and discriminant analysis integrated with variable selection and cross validation; leave-one-out procedure were used to the multivariate statistical analysis. 73 patients' demographics and clinical features were included in the analyses.

Results: The performance of the logistic regression integrated backward variable selection and leave-one-out is found higher than other logistic regression and discriminant in terms of accuracy (0.863). Glasgow Coma Score, fever, diabetes, spo2, pulse, WBC and CRP are found as the most important factors classifying the ICU admission.

Conclusion: As a result of this study, especially having a fever, high pulse and CRP increase intensive care admissions whereas high spo2 and WBC decrease intensive care admissions. Also, this study can be improved by using other classification methods such as machine learning techniques and the sample size can be expanded.

Keywords: COVID-19, Intensive Care Unit (ICU) Admission, Clinical Prognosis, Logistic Regression, Discriminant Analysis

References

Y. Chen et al., "An Interpretable Machine Learning Framework for Accurate Severe vs Non-Severe COVID-19 Clinical Type Classification," SSRN Electron. J., Oct. 2020, doi: 10.2139/ssrn.3638427.

D. Assaf et al., "Utilization of machine-learning models to accurately predict the risk for critical COVID-19," Intern. Emerg. Med., vol. 15, no. 8, pp. 1435–1443, Nov. 2020, doi: 10.1007/s11739-020-02475-0.

D. Liao et al., "Haematological characteristics and risk factors in the classification and prognosis evaluation of COVID-19: a retrospective cohort study," Lancet Haematol., vol. 7, no. 9, pp. e671–e678, Sep. 2020, doi: 10.1016/S2352-3026(20)30217-9.

J. Gong et al., "A tool for early prediction of severe coronavirus disease 2019 (covid-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China," Clin. Infect. Dis., vol. 71, no. 15, pp. 833–840, 2020, doi: 10.1093/cid/ciaa443.

J. Wu et al., "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," medRxiv,2020.

WHO, "WHO Director-General's Opening Remarks at The Media Briefing on COVID-19," Press Brief., 2020.

S15

Semiparametric shared frailty models for right-censored data using an EM algorithm

Aysun CETINYUREK YAVUZ[1], Philippe LAMBERT[2]

[1] *Danone Nutricia Research, Utrecht, Belgium*
[2] *Institut des Sciences Humaines et Sociales, Université de Liège, Liège, Belgium*
[2] *Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

Abstract

Aim: In survival analysis frailty models are used to accommodate unobserved heterogeneity. Clustered time-to-event data is handled by inclusion of a latent frailty term. The latent frailty is assigned a parametric distribution, typically, a gamma distribution due to its conjugacy in the Cox proportional hazards model. Due to the limitations of a gamma distribution, several other frailty models are also considered in the literature, e.g. log-normal, positive stable, etc. However the conjugacy property is lost when other distributions are used. Besides, each distribution imposes a different correlation structure.

Methodology: This makes the use of more flexible distributions attractive as they can approximate many different forms. In this paper, we propose to use gamma shape mixtures (GSM) for modelling frailty. Expectation Maximization (EM) algorithm facilitates the use of more complex frailty distributions, such as GSM. Here, we present a shared frailty proportional hazards model with a flexible form for baseline log-hazard using P-splines and the gamma and GSM distribution for the frailty.

Findings: The methodology is illustrated on a dataset. The dataset is also analysed using frailtyEM and survival packages in R, and the results are compared with those of the presented frailty models.

Result: The combination of the EM algorithm and of a GSM distribution provides a more flexible modelling framework for analysing clustered right-censored data.

Keywords: flexible; right-censored; shared frailty; proportional hazards; P-splines, misspecification.

S16

Gradyan Artırma ve Hafif Gradyan Artırma Algoritmaları ile Kırım Kongo Verisi Üzerine Uygulama

Osman DEMİR[1], Yunus Emre KUYUCU[1]

[1] *Tokat Gaziosmanpaşa Üniversitesi Tıp Fakültesi Biyoistatistik AD. TOKAT*

Amaç: Makine öğrenimindeki boosting (artırma) topluluk modellerinden olan gradyan artırma (GBM) ve light gradyan artırma (LightGBM) yöntemlerinin Kırım Kongo Kanamalı Ateşi hastalığında ölüm tahmininde kullanılması amaçlanmaktadır. Ayrıca uygulanan bu sınıflayıcıların performansları karşılaştırılmaktadır.

Materyal ve Yöntem: Veri seti, 209 Kırım Kongo Kanamalı Ateşi (KKKA) teşhisi almış hastanın sağkalımı üzerine yapılan çalışmaya aittir. Veri setinde ölen hasta sayısı 28 (%13,4) ve sağ kalan 181 (86,6) kişidir. Ölen hastaların tahmin edilmesinde topluluk öğrenme yöntemlerinden boosting (artırma) yöntemi uygulanmaktadır. Sınıflayıcı olarak da gradyan artırma (GBM) ve light (hafif) gradyan artırma (lightgbm) yöntemleri kullanılmaktadır.

Bulgular: Ölen hastaların tahmin edilmesinde gbm sınıflayıcısına ilişkin doğruluk değeri 0,967 iken lightgbm sınıflayıcısında 0,918 olarak bulunmuştur. Duyarlık değeri gbm için 1,000 iken lightgbm için 0,962 olarak elde edilmiştir. Eğri altı değerleri ise gbm ve lightgbm için sırasıyla 0,932 ve 0,892 olarak bulunmuştur. Çalışmada kullanılan veri setine göre, performans ölçüleri karşılaştırıldığında gbm algoritması daha yüksek performans göstermektedir.

Sonuç: Topluluk öğrenme yöntemlerinden, gbm, lightgbm sınıflayıcısına göre genel olarak daha başarılı bir sınıflama gerçekleştirmiştir.

Anahtar Sözcükler: Makine öğrenme, topluluk öğrenme, sınıflayıcılar, gbm, lightgbm, Kırım Kongo Kanamalı Ateşi.

Kaynaklar

1. Greenwell, B., et al., Package ‘gbm’. R package version, 2019. 2(5).

2. Aktas, T., et al., Mean platelet volume (MPV): a new predictor of pulmonary findings and survival in CCHF patients? 2017.

3. Ke, G., et al., Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017. 30: p. 3146-3154.

4. Chen, T., et al., Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. IEEE Access, 2019. 7: p. 150960-150968.

5. Freund, Y., & Schapire, R. E. (1996). Schapire R: Experiments with a new boosting algorithm. In *in: Thirteenth International Conference on ML*.

6. Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*(2), 197-227.

S17

A Novel Approach to Clustering RNA-Sequencing Data Based on Self-Organizing Maps Algorithm

Ahu CEPHE[1,2], *Necla KOÇHAN*[3], *Gözde ERTÜRK ZARARSIZ*[2,4], *Vahap ELDEM*[5], *Erdem KARABULUT*[6], *Gökmen ZARARSIZ*[1,2,4]

[1]*Erciyes University, Rectorate, Institutional Data Management and Analytics Unit, Kayseri, Turkey*
[2]*Erciyes University, Drug Application and Research Center (ERFARMA), Kayseri, Turkey*
[3]*İzmir Ekonomi University, Faculty of Arts and Sciences, Department of Mathematics, İzmir, Turkey*
[4]*Erciyes University, Faculty of Medicine, Department of Biostatistics, Kayseri, Turkey*
[5]*İstanbul University, Faculty of Science, Department of Biology, İstanbul, Turkey*
[6]*Hacettepe University, Faculty of Medicine, Department of Biostatistics, Ankara, Turkey*

ahucephe@erciyes.edu.tr

Summary

Using gene expression data, subtypes of diseases can be determined by clustering analyses and personalized treatments can be applied for these subtypes. Since some disadvantages of microarray technology, which has been used for many years for these analyses, researchers have started to use RNA-sequencing technology. However, performing clustering analyses using high-dimensional RNA-sequencing technology, which can measure the expression level of tens of thousands of transcripts simultaneously, is not easy due to the problem of over-dispersion in RNA-sequencing data. When the literature is examined, there are various clustering approaches that identify new disease subtypes using RNA-sequencing data (1-3). These approaches, most of which are based on discrete probability distributions, do not deal with the problem of over-dispersion in RNA-sequencing data. For this reason, it has been observed that in clustering analyses using the voom (variance modeling at observational level) transformation method, which is based on the logic of modeling the mean-variance relationship instead of determining discrete distributions, more accurate results are achieved than approaches using discrete distributions (4-5). It has been shown in previous studies that the self-organizing maps (SOM) clustering method, which processes big data effectively, is robust to data noise, and is used in high-dimensional data visualization, gives very good results in gene expression data (6). SOM clustering algorithm, an artificial neural network approach, creates a projection that preserves the adjacency relationship in the data, unlike clustering algorithms that group by making maximum distance between points in the cluster and minimum of the distances between any two points, one from each cluster (7). This method is also robust to noise, called outliers, which is also abundant in RNA-sequencing data. This study aims to introduce a new approach that can be used to detect of subtypes of diseases by integrating the voom transform and SOM algorithm, which are powerful methods in the literature. The experimental results performed in the study showed that the SOM algorithm used together with the voom method gave the best results compared to the existing methods.

Keywords: Clustering, self-organizing maps, RNA-sequencing, voom, next-generation sequencing

References

1. Witten, D.M. (2011). Classification and clustering of sequencing data using a poisson model. Ann. Appl. Stat. 5(4), 2493–2518.

2. Si, Y., Liu, P., Li, P., Brutnell, T. (2014). Model-based clustering of RNA-seq data. Bioinformatics 30(2), 197–205.

3. Robinson, M.D., Mccarthy, D.J. and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 139–140.

4. Cephe, A. (2019). Rna-Dizileme Verilerinin Kümelenmesinde Yeni İstatistiksel Yaklaşımlar. Erciyes Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Ana Bilim Dalı Yükseklisans Tezi.

5. Law C.W., Chen Y., Shi W., Smyth G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol, 15:29.

6. Nan, F., Li, Y., Jia, X.Y., Dong, L., Chen, Y. (2019). Application of improved SOM network in gene data cluster analysis[J]. Measurement, 145.

7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T. (1999). Interpreting gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences, USA, 96:2907-2912.

S18

Feature Selection and Classification in Lung, Prostate and Breast Cancer Microarray Gene Expression Data

Özlem ARIK[1], Erdem KARABULUT[2]

[1] *Kutahya Health Sciences University, Faculty of Medicine, Department of Biostatistics, Kütahya*
[2]*Hacettepe University, Faculty of Medicine, Department of Biostatistics, Ankara*

Abstract

High number of feature (p) and low sample size (n) parameters in microarray gene expression data obtained through DNA microarray technology adversely affect the performance measure values of machine learning algorithms. For this reason, it is important to apply data mining and feature selection methods in modeling microarray gene expression data. In data mining, the data is separated by using the classification methods and the common features of the data, and it is decided in which class the new data will be added. The main purpose of this study; it is to obtain classification models and measure their performance after making feature selection on the datasets of three different cancer types obtained from the NCBI-GEO database. Feature (gene) selection was performed with varFilter, nsFilter, lasso and limma feature selection methods on microarray gene expression data of lung, prostate and breast cancers. Classification models were obtained through support vector machines, k-nearest neighbor and deep learning methods in datasets with feature selection. In the study conducted with the R program, the performance values of the classification models created after feature selection with the limma method among the feature selection methods were found to be higher. Among the obtained classification models, the classification success of deep learning was better than other classification methods.

Keywords: *Cancer, Microarray, Limma, Deep Learning, Performance*

Introduction

The importance of bioinformatics is increasing day by day, as well as biostatistics, which is used in collecting and analyzing relevant data in research in the field of health and making the right decisions through the results (1, 2, 3). Bioinformatics; it includes the fields of biology, computers, mathematics, statistics and genetics and one of the most important research topics is gene analysis. Thanks to the DNA microarray technology used in this field, known and unknown functions of genes are detected (4, 5, 6). Thanks to the developments in genome sequencing and bioinformatics, better diagnosis, treatment and prevention studies are carried out by understanding the genome structure of cancer cells and the changes in the internal dynamics of cancer (7). Finding genes directly related to the disease with microarray gene expression data is important in the diagnosis and classification of diseases such as cancer. Since genetic data is very large, analysis can be made with data mining methods and computer programming (8, 9). Data mining, which first entered the literature in the 1980s, is widely used today (10-12). Data mining, which is an interdisciplinary field, is becoming more and more accepted day by day with the combination of concepts such as statistics, machine learning and artificial intelligence, as classical statistical methods cannot provide valid and reliable results for large amounts of data (13, 14). The methods used in data mining are generally divided into two as predictive and descriptive (15, 16). Predictive methods are used to obtain the results of new situations encountered in relation to the problem, thanks to the model created by the data of known results of any event. Classification is also among the predictive methods. Thanks to the classification methods, a classification model is obtained with the data known to belong to which class, and it is decided to which class the newly added data will be included (14).

The term bioinformatics has started to be used after the mid-1980s. The need for bioinformatics has increased for the processing of the genetic information generated as a result of the Human Genome Project. Therefore, the type of data that bioinformatics studies mostly is genetic data and accordingly gene expression data (5, 17-19). In the second half of the 20th century, with the increase in biological knowledge, large-scale biological data are organized, analyzed and made more understandable thanks to bioinformatics, which is an interdisciplinary science. NCBI (National Center for Biotechnology Information), which is open to the use of researchers in bioinformatics and was established in Maryland in 1988 to store, organize and use nucleotide sequence information, is the most important web-based biological database (5, 20).

The main material of genetic studies, which started with Mendel's studies in the 1800s, is DNA (Deoxyribo Nucleic Acid). DNA is found in all living things with cells, it carries the biological information needed for the development of the living thing and takes part in the transfer of this information to the next generations (21-23). The gene, which forms the root of the word genetics, is the DNA regions located in the chromosomes in the cell nucleus, which have genetic tasks such as defining physical characteristics, and have start and end points (24). Gene expression, which shows whether the genes are active or not and how active they are if they are active for a situation being studied, is the conversion of genes from DNA to RNA structures and proteins. There is a positive linear relationship between overproduction of protein and high level of gene expression (25). All of our organs contain the same genetic material. However, because genes are expressed differently in different cells, cells such as breast, lung and brain do not have the same functions (26, 27). Thanks to the microarray technology, which is one of the steps taken with the advancing technology, the expressions of the genome of an organism can be examined in one go (28, 29). Characteristics such as being fast, examining the activity of genes in sick and healthy cells, and being able to categorize diseases are the advantages of microarray technology. Due to the size and complexity of the data obtained, the need for computational genomic approaches for analysis and interpretation has increased. Biostatistical analysis and bioinformatics also have a great place in meeting these needs (6, 30).

Microarrays are also referred to as chips consisting of thousands of spots where thousands of different DNA fragments are synthesized or inserted. A solid surface made of glass, plastic, or silicon is called a chip. The probe is each point on the surface of the chip (29, 31). If the gene is not expressed or read, the probe will appear black. Green, healthy people; red indicates sick people. If the state of being sick or healthy is close to each other, it is shown in yellow. By computer analysis, these colors are converted into numerical values and made available for analysis (32, 33). The most important of the publicly available microarray databases is; it is the GEO (Gene Expression Omnibus) of American origin and under NCBI, the world's most comprehensive biological database. The other is ArrayExpress under EBI, a large and comprehensive biological database of European origin (34). The real data sets used in this study also belong to the NCBI-GEO database. Gene expression data matrix of size mxn with genes in the row and samples in the column is in the form of a row data structure (9, 35). In order to perform analyzes such as feature selection and classification, the gene expression data matrix is transposed and samples are placed in the rows and genes are placed in the columns. With the application of data mining and statistical methods on the obtained gene expression data matrix, genes that affect diseases such as cancer can be determined, genes with common functions can be clustered, and individuals can be classified as sick-healthy (25). Some changes in genes that control how cells grow and divide, that is, how they function, cause cancer (23). Early diagnosis is very important in cancer, and working with gene expression data is of great importance in the diagnosis and classification phase (36).

In this study, large-scale microarray gene expression data for lung, prostate and breast cancers were used, with features (genes) and response variable (structure of the tumor) in the column, and individuals in the row. After the data preprocessing steps, important and meaningful features were selected with the varFilter, nsFilter, lasso (least absolute shrinkage and selection operator) and limma (linear models for microarray data) feature selection methods. Then, in order to create models of sick-healthy classification with the selected data, support vector machines and k-nearest neighbor classification methods, which are frequently used in data mining, were preferred together with deep learning. Model performance measures such as accuracy, sensitivity, specificity and area under the ROC curve (AUC) were used to evaluate the performance of classification models obtained by using feature selection methods and classification methods on microarray gene expression data of three different cancer types.

Method

There is a similar definition of the concept of a feature and a variable that expresses properties or situations that take different values from sample to sample. While the number of samples is low in microarray gene expression data, which is generally among high-dimensional data; the number of feature (gene) is quite high. Thus, it is a problem to work with microarray gene expression data which has unwanted data structure (37). For the purpose of the application planned to be made on a large data set such as microarray gene expression data, the process of determining the best feature subset that can represent the original data set by removing unnecessary features instead of using all of the features is called feature selection. Thanks to feature selection, better models are obtained in terms of speed and success performance (38, 39).

There are different methods used to perform feature selection. In general, these methods are divided into three as statistical methods, wrapper methods and embedded methods (40). Statistical methods that make selections using only statistical information are also called filtering methods. Methods such as heuristic search, genetic algorithm, particle swarm optimization are among the wrapper methods. In these methods, data mining methods are used as a tool for feature selection. Embedded methods are the methods in which the feature selection method and the data mining method are applied simultaneously (38, 41). Especially in microarray gene expression data, the use of embedded methods is more ideal (42). Since there are many features in the data sets used in this study, four different feature selection methods in the R program were used in order to find important and meaningful features that will affect the classification result more. When computer memory does not allow analysis in high-dimensional data sets such as microarray gene expression data, there are R packages available for performing applications such as feature selection, classification, clustering. It is suitable for integrating the R program with the Bioconductor environment, which includes genomic data sources and open source analysis tools, especially numerous microarray data sets (43). Gene expression data of microarray studies are contained in the Bioconductor ExpressionSet object. An ExpressionSet object has been created so that different information sources can be converted into a single structure and become more useful (43, 44). In this study, microarray gene expression data of cancer types taken from the NCBI-GEO database were converted into ExpressionSet object via Bioconductor and made ready for application. ***geneFilter*** and ***CMA*** packages are used to perform feature selection on microarray gene expression data using the ExpressionSet object (45). The varFilter() and nsFilter() functions in the ***genefilter*** package, which is one of the R packages created to eliminate computer-related problems such as speed and memory, are used to make feature selection on high-dimensional data sets.

In varFilter method; variance values are obtained for each of the features in the data set. Among the variance values ordered from the largest to the smallest, those that come before a certain limit are determined. Features with determined variance values are selected for use in later stages (43). In other words, while selecting features that vary a lot between samples as a result of selection with varFilter, features that show little change are discarded. The var.cutoff value in the varFilter() function is used to express how much of the total features in the data set is desired to be studied. For example, var.cutoff=0.80 to select 20% of the features and var.cutoff=0.90 to select 10% of the features (45).

As in varFilter, the nsFilter method is used for feature selection, if the information in the annotation package of the data set converted to the ExpressionSet object, that is, the explanation information is to be selected. By using the nsFilter() function, features that show low signal and with variance calculation, features that do not vary much between samples, are not selected (43, 46). Using a criterion such as the var.cutoff value in the varFilter method, it is not determined in advance how much of the features will be studied.

In order to create a model with good performance, taking into account the data set, one of the embedded methods that makes use of the search algorithm is lasso (least absolute shrinkage and selection operator) (41). Lasso was first developed by Tibshirani in 1996 as a method that can perform coefficient estimation and variable selection simultaneously in regression analysis (47, 48). With regression analysis, the value of the dependent (response) variable is estimated by using the value of the independent variable(s). The least squares method is used to construct the linear regression model, which is used in cases where there is a linear relationship between the independent variable(s) and the response variable. In the method, the coefficients related to the independent variables, namely the parameters, are estimated. However, when there are too many independent variables, some problems arise, such as multicollinearity, in which a linear or close to linear relationship is observed between the variables (49). When there is multicollinearity, the coefficient estimates are uncertain and the variances and standard errors of the estimates become larger, and R^2 becomes larger than it should be. Different methods are used by making some changes in the regression model to be created to estimate the value of the response variable. Lasso is one of these methods (41). In gene expression data, independent variables correspond to features, and there are many features. In the model created by using the features in the data set with the Lasso method, problems such as overfitting and multicollinearity are eliminated, and the coefficients of less important features are calculated as zero. Thus, feature selection is made automatically with lasso (36). For the application of the Lasso method, the ***glmnet*** package was used together with the ***CMA*** package.

Limma (linear models for microarray data), one of the methods in the geneSelection()function in the ***CMA*** package of Bioconductor in the R program, was first introduced by Smyth in 2003 (50, 51). In the analysis of gene

expression data obtained through RNA sequencing or microarray technologies, limma is used to identify differential gene expressions and analyze experimental designs (52, 53). In this study, the application of the limma method on datasets that have two classes as features and sick-healthy, is examined whether there is a difference between the two groups as sick-healthy in terms of features (genes). Feature selection is performed by identifying differentially expressed genes between the two groups (54). The ***limma*** package was also used to implement the limma method used in the study (52).

Classification methods are used to obtain a meaningful model for predicting the future by using data including independent variables and categorical response variable (10). In this study, while the features express the independent variables, the response variable consists of two classes as sick-healthy. In the microarray gene expression data, which includes samples, features and categorical response variable, it is known what class the samples are in. With the classification models created using this information, when a new sample arrives, it is predicted in which class this sample will be (14, 55). There is a process followed for the classification process. After the data preprocessing step, not all of the available datasets are used for model building. In the training dataset, a model with known class examples and classification rules is created. In the test dataset, the obtained model is tested and its accuracy is measured (25, 56). In this study, 75% of the data set was used as training and 25% as a test dataset, and 5-fold cross validation was performed for model generalization. In the cross validation method, the dataset is divided into k subsets. k-1 clusters are used in training and 1 in testing. The process is repeated k times and the accuracy values obtained each time are averaged and the accuracy performance of the model is calculated (13, 43). When working with large data sets such as microarray gene expression data, classification methods, one of the most frequently used methods of data mining, are applied after selecting effective and meaningful features in order to both reduce the computation time and obtain models with better performance. When we look at the studies involving gene expression data, support vector machines and deep learning method were used in addition to k-nearest neighbors, which are among the most preferred classification methods (40).

The use of the Support Vector Machines (SVM) classification method, which was first revealed by Vladimir Vapnik and Alexey Chervoenkis in the 1960s, became widespread with the first successful applications in the 1990s (35, 57). Although it was developed for situations where the number of classes is two, over time it has been expanded to be applicable for data sets where the number of classes is more than two and cannot be separated linearly (58). It is a very preferred method because it gives classification results with high performance in large data sets such as gene expression data with a large number of features and in many other fields (35, 58). In the SVM classification method, when distinguishing between instances of classes by using feature sets belonging to different classes, the plane that will make the correct classification is determined and called the hyperplane. Points that make the boundary width limited are also called support vectors. The farther the borders are from each other, the better (37, 40, 55). In this study, the train() function of the ***caret*** package is used to obtain the SVM classification model.

The k-Nearest Neighbor (kNN) classification method was first introduced by Fix and Hodges in the early 1950s and developed and popularized by Cover and Hart towards the end of the 1960s (38-40). In the kNN classification method, which is the most basic of the sample or memory-based methods, the class of the sample, which is unknown to which class, is determined by pre-classifying through the training dataset (41, 42). The k neighbors of the sample to be classified are determined in the training set with the closest distance, and the new sample is included in that class in whichever class most of these neighbors are. For the classification process, using distance measures, the similarity of the new sample between the samples in the training dataset is checked, and the class prediction is made for the sample by determining the closest training dataset sample (40). For this, distance measures such as Minkowski, Euclid, Manhattan are used. In previous studies, the Euclidean distance measure was generally used (43). To obtain the Euclidean distance measure, the square root of the sum of the squares of the difference between the features is taken. There are different values for each of the classes, and the class with the highest value is assigned the newcomer, that is, the tested instance (40, 44). Therefore, it is important that the features of the classes are determined (43). In the study, the knnCMA() function of the ***CMA*** package was used to obtain the kNN classification model.

In the 2000s, Geoffrey Hinton and Ruslan Salakhutdinas showed how multilayer artificial neural networks work. In the Deep Belief Network study in 2006, the way multi-layered deep structures work and the self-completion of missing features were expressed and this was called Deep Learning (DL) (45). DL, of which many methods were

put forward, was used by scientists for the first time in 2012 and over time, it has gained a widespread use (46). In the DL method, which is an advanced extended approach of artificial neural networks consisting of three layers, input, hidden and output, there is more than one hidden layer. Thus, the data can be comprehensively represented (45). Thanks to the hidden layers, the result in the output layer is obtained. DL methods generally cover deep structures of successive layers. By successive layer, the output of each layer forms the input of the next layer. Thus, learning is provided in a hierarchical manner. Using the neurons in the input layer, the neurons in the hidden layers are calculated with the linear function y=f(x,w). The neurons in the output layer are obtained by applying the activation functions on the calculated neurons. The output layer obtained as a result of the operations is a nonlinear form of the input layer. Nonlinear problems are generally tried to be solved with DL methods because it gave better results in these types of problems than other methods (58). In RNA sequencing and microarray gene expression data, there are fewer studies in which the DL method is used for classification compared to the studies in which classical data mining classification methods are used (46, 52, 58). As in other classification methods, the data set is divided into two as training and test datasets. The ***h2o*** package developed by the H2O.ai team (2017) is then used to create the classification model. Datasets are converted to h2o objects by running this package with the h2o.init() function. By means of the h2o.deeplearning() function, the DL model is obtained from the h2o data object (43). In this study, in order to show the effect of feature selection methods on the performance of classification methods and to make comparisons, in order to ensure that the results are at a certain standard, DL classification models are obtained by applying feature selection methods as in other classification methods (52).Especially in the field of health, cancer diagnosis and classification, drug development, medical image and signal data, DL methods are frequently applied. It is increasingly preferred as a result of its high performance in terms of accuracy in solving questions. (46).

Various model performance measures are used to evaluate and interpret the performance of classification methods regarding how accurately they classify. Since there is no imbalance in the distribution of the number of classes in the data sets discussed in this study, the most preferred measures such as accuracy, sensitivity, specificity and AUC were used. A classification table is used for the calculation of these measurements. There are four possible outcomes in a dataset where the number of classes is two. These results are as in the classification table given in Table 1. (47,58).

Table 1. Classification table of real and predicted results

Predicted Class	Real Class	
	Positive	Negative
Positive	A (True Positive- TP)	B (False Positive - FP)
Negative	C (False Negative - FN)	D (True Negative - TN)

The performance measures used in the study can also be calculated with the values in the classification table. Accuracy is a very preferred simple method in determining the performance of the classification method and is a general success measure. It is the ratio of the number of correctly classified samples (TP+TN) to the total number of samples (TP+FP+FN+TN). Sensitivity is the proportion of samples whose predicted class is positive among samples that are actually in the positive class. It is the performance of the classification method in identifying samples with positive values. Specificity is the proportion of samples with a predicted negative class among samples that are actually in a negative class. It is the performance of the classification method in identifying samples with negative values. AUC, which was first used in signal detection in the 1950s, is widely used in biomedical studies. The main purpose of constructing the ROC curve is to examine the result obtained from the classification method in terms of accuracy values. Therefore, firstly, the sensitivity and specificity values are calculated. AUC obtained on the ROC graph, which consists of 1-specificity on the horizontal axis and sensitivity values on the vertical axis, takes values between 0.5 and 1. The closer the AUC is to 1, the better the classification performance of the method used; The method with an AUC of 0.5 is quite unsuccessful in classification (48, 58).

The response variable of the data used in this study consists of two classes as sick-healthy. The power of the classification method used to discriminate between sick and healthy individuals is shown by AUC (49).

The real datasets used in the study consist of gene expression data obtained from microarray experiments of three different cancer types: lung, prostate and breast. Data sets can be accessed through the GEO data repository offered by NCBI. The GEO data repository can be accessed at www.ncbi.nlm.nih.gov/geo/ (25, 43). In GEO, the GDS dataset is a set containing information about the platform and sample data records. Preprocessing steps in microarrays such as normalization and background processing have been done on the data set, and it is assumed that the measurements are made in an equivalent way. Each dataset has code starting with GDS. While there are 22283 features in the prostate and breast cancer datasets from the microarray gene expression data used in this study; lung cancer dataset has 54675 features. In the considered data sets, there is a response variable that includes two class values as sick and healthy. Considering the total number of samples, there are 120 samples (60 patients-60 healthy) in the lung cancer dataset, 79 samples (40 patients-39 healthy) in the prostate cancer dataset, and 42 samples (24 patients-18 healthy) in the breast cancer dataset. The sample numbers of the sick-healthy classes in each data set are equal or very close to each other. Data preprocessing of genomic or clinical datasets is important in data analysis. Especially in large data sets, if the data preprocessing steps are taken into consideration before starting the analysis of the data, the analysis is carried out without any problems. n the study, missing value analysis, data scaling, imbalance in the distribution of class numbers, and outlier analysis were performed in microarray gene expression data of three different cancer types (43). It has been concluded that there are no missing and outlier values in the datasets considered, and that unbalanced datasets are not observed in the distribution of class numbers. By using Z-score scaling, which takes into account the mean and standard deviation of the data, the superiority of the features over each other is prevented. In addition, while applying the feature selection and classification methods; Since the data set should be in matrix form with rows samples and columns containing features, the operations were continued by taking the transpose of the gene expression data matrix, which is in the form of row data structure.

Results

In this section, the findings of the application of the discussed feature selection and classification methods in microarray gene expression data of lung, prostate and breast cancers are given in Table 2, Table 3, Table 4, Figure 1, Figure 2 and Figure 3. First, the results of lung cancer are given in Table 2.

Table 2. Comparison of the performances of classification models created by using features determined by feature selection methods in the lung cancer dataset.

Feature Selection Methods	**Classification Methods**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
varFilter	SVM	0,833	0,928	0,751	0,839
	kNN	0,733	0,562	0,929	0,753
	DL	0,951	1,000	0,923	0,945
nsFilter	SVM	0,900	0,882	0,923	0,889
	kNN	0,767	0,562	1,000	0,751
	DL	0,933	1,000	0,926	0,950
lasso	SVM	0,958	0,967	0,950	0,960
	kNN	0,960	0,970	0,950	0,960
	DL	0,965	0,955	0,970	0,961
limma	SVM	0,958	0,983	0,933	0,950
	kNN	0,970	0,970	0,970	0,962
	DL	0,986	1,000	0,975	0,988

When the results of the lung cancer data set are examined, the performance of the model obtained by the DL classification method is better than the other methods after applying the varFilter feature selection method. The second best performing model is SVM. In the performance of the classification model obtained with kNN, the performance measure values are quite low compared to other models, except for the specificity value.

Among the classification models created by using the nsFilter feature selection method, the best performance is obtained with the DL method. The success of the DL model is followed by the SVM. The best model in terms of specificity performance is kNN. However, in other performance measure values, the performance of kNN is generally lower than other methods.

In the lasso feature selection method, the performance values of SVM, kNN and DL are approximately close to each other and are quite good.

Finally, among the classification models obtained using the limma feature selection method, the best performance is obtained with the DL method. After DL, the performance of kNN come. The success of the classification model obtained with SVM is close to kNN. The results obtained are given in Figure 1. by means of graphics.

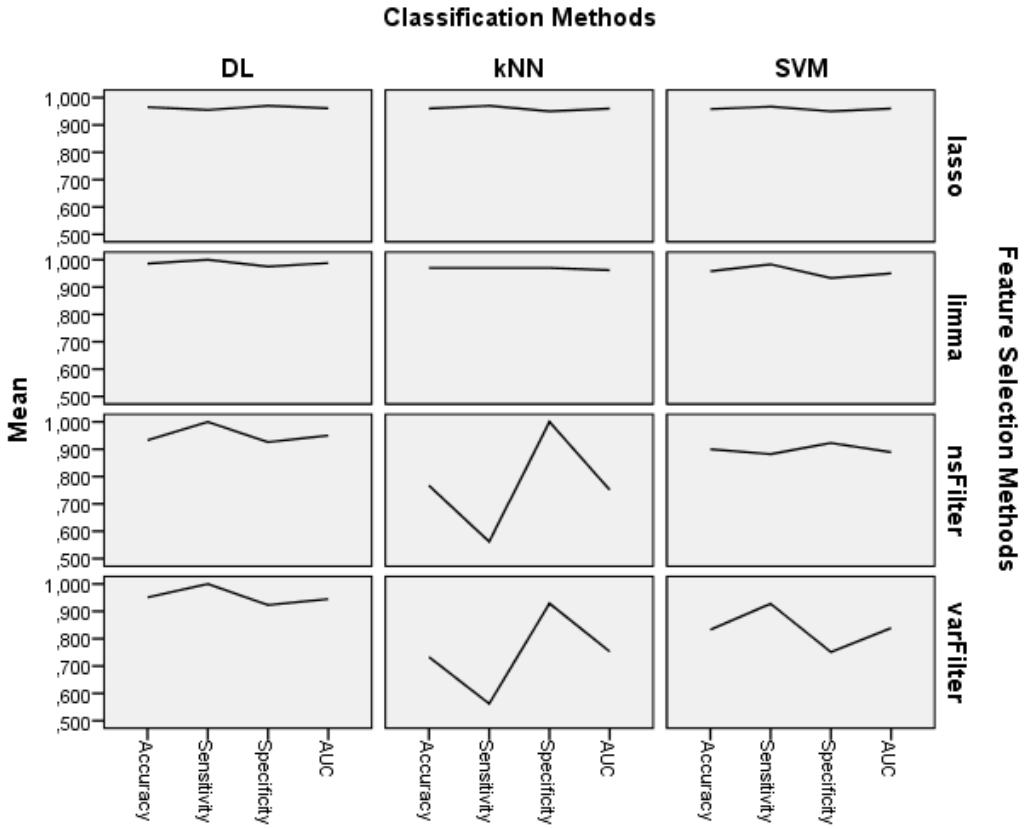


Figure 1. Comparison of accuracy, sensitivity, specificity and AUC performances of classification methods in different feature selection methods for lung cancer dataset.

While the success of the classification models obtained by the lasso and limma feature selection methods are higher; In the varFilter and nsFilter feature selection methods, the classification methods have lower performance. DL performs better within classification methods.

Table 3. Comparison of the performances of classification models created by using features determined by feature selection methods in the prostate cancer dataset.

Feature Selection Methods	**Classification Methods**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
	SVM	0,600	0,666	0,545	0,606
	kNN	0,600	0,666	0,545	0,606
varFilter	DL	0,845	0,810	0,990	0,858
	SVM	0,800	0,666	0,909	0,809
	kNN	0,650	0,540	0,860	0,660
nsFilter	DL	0,900	1,000	0,870	0,888
	SVM	0,578	0,511	0,625	0,598
	kNN	0,581	0,614	0,550	0,604
lasso	DL	0,870	1,000	0,850	0,875
	SVM	0,581	0,561	0,600	0,585
	kNN	0,580	0,664	0,500	0,650
limma	DL	0,950	1,000	0,900	0,960

When the results of the prostate cancer dataset are examined in Table 3., the performance of the model obtained by the DL classification method after applying the varFilter feature selection method is much better than the other methods. The performance values of the classification models obtained by using the SVM and kNN methods are the same.

In the nsFilter feature selection method, the success of the DL model is followed by SVM. On the other hand, kNN is the classification method with the lowest performance in the nsFilter feature selection method in the prostate cancer dataset.

In the lasso feature selection method, the best performance measure values are obtained with the DL classification method. The performance measure values of the classification models created by kNN and SVM methods are close to each other and come after DL.

Finally, among the classification models obtained using the limma feature selection method, the best performance is obtained with the DL method. Classification models with lower performance are obtained with the SVM and kNN methods, whose performance measure values are close to each other. The results obtained are given in Figure 2. by means of graphics.

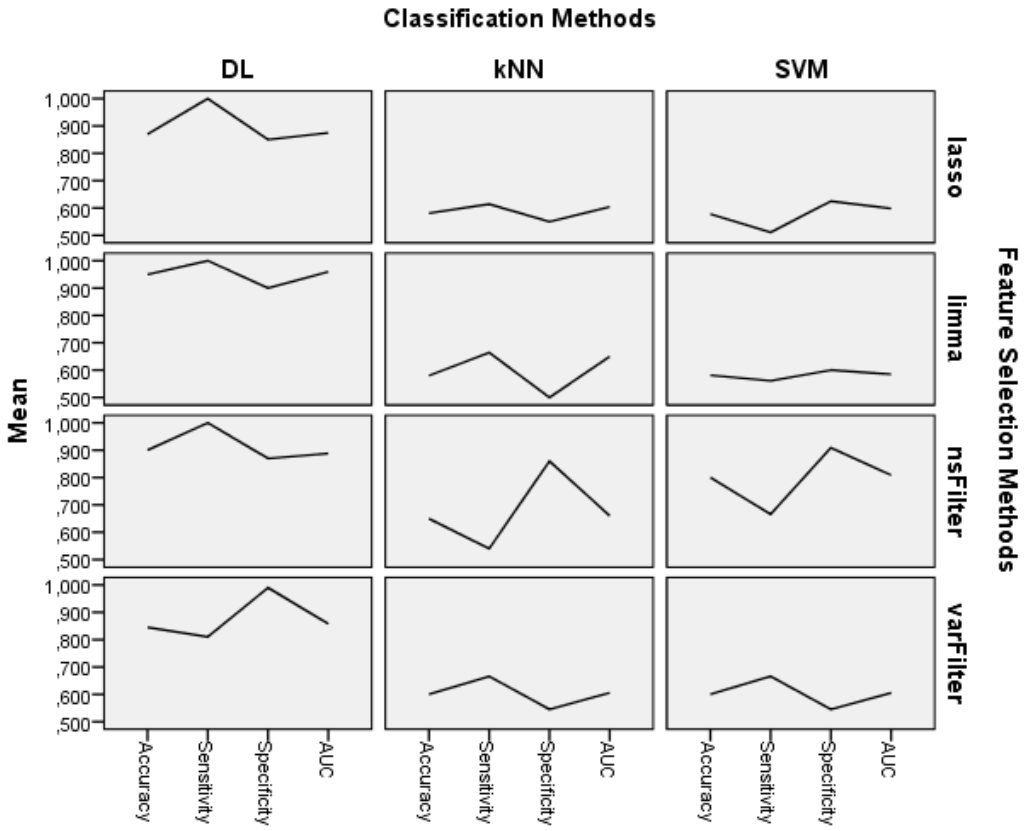


Figure 2. Comparison of accuracy, sensitivity, specificity and AUC performances of classification methods in different feature selection methods for prostate cancer dataset.

In the prostate cancer dataset, DL has a higher performance in general among the models obtained with the classification methods. Especially in limma and nsFilter feature selection methods, the success of DL is better.

Table 4. Comparison of the performances of classification models created by using features determined by feature selection methods in the breast cancer dataset.

Feature Selection Methods	Classification Methods	Accuracy	Sensitivity	Specificity	AUC
varFilter	SVM	0,540	0,333	0,750	0,541
	kNN	0,727	0,875	0,673	0,691
	DL	0,797	0,840	0,790	0,800
nsFilter	SVM	0,540	0,333	0,750	0,541
	kNN	0,818	0,963	0,667	0,798
	DL	0,818	0,960	0,800	0,810
lasso	SVM	0,648	0,790	0,450	0,640
	kNN	0,639	0,740	0,500	0,632
	DL	0,909	1,000	0,850	0,933
limma	SVM	0,671	0,710	0,633	0,667
	kNN	0,633	0,690	0,567	0,630
	DL	1,000	1,000	1,000	1,000

When Table 4 of the breast cancer dataset is examined, the performance of the model obtained by the DL classification method after applying the varFilter feature selection method is better than the other methods. The performance values of kNN come after DL. The performance values of the classification model obtained by SVM, where the sensitivity value is low, are generally the lowest among the classification models obtained with other methods.

The results of the performance of the classification models created by using the nsFilter feature selection method are similar to varFilter. The best performance is obtained with the DL method. The performance of the classification model obtained by the kNN method comes after the DL. Compared to other classification methods, SVM has lower performance values.

Among the classification models obtained using the lasso and limma feature selection method, the best performance is obtained with the DL method. The performance values of the classification models obtained by SVM and kNN methods are close to each other and they have lower performance than DL.

The results obtained are given in Figure 3. by means of graphics. Looking at Figure 3, DS generally has a higher performance among the models obtained by classification methods in the breast cancer dataset. The success of the classification models obtained by using mostly lasso and limma feature selection methods are higher.

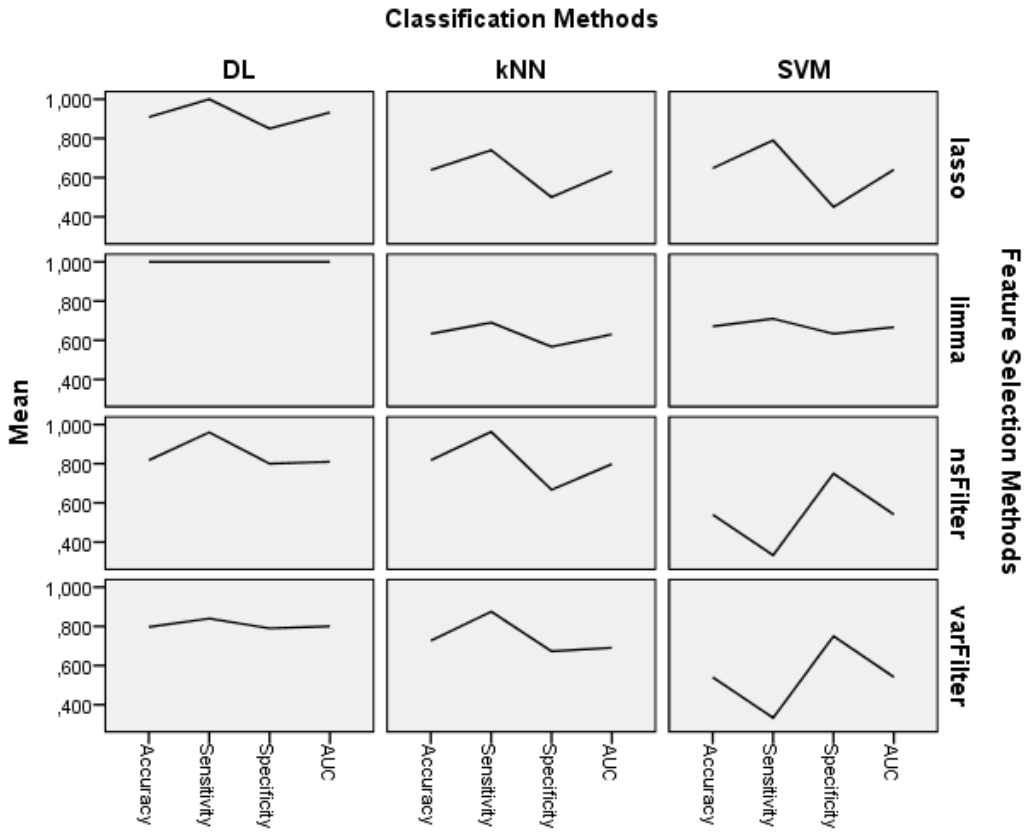


Figure 2. Comparison of accuracy, sensitivity, specificity and AUC performances of classification methods in different feature selection methods for breast cancer dataset.

Discussion and Conclusions

It is aimed to show the effect of feature selection methods used to reduce the number of features in microarray gene expression data, which has data type with high number of features and low number of samples, on the performance level of classification methods frequently used in data mining. For this purpose, three real data sets were studied. SVM, kNN and DL classification methods were applied to the data sets containing important features obtained by the application of varFilter, nsFilter, lasso and limma feature selection methods on microarray gene

expression data of three different cancer types, and classification models in which the classification was made as sick and healthy were created. Model performance measure values in the form of accuracy, sensitivity, specificity and AUC for these models were obtained. When the results of each data set are examined in the study conducted through the R program; better performance measure values are obtained in lasso and limma feature selection methods in lung and breast cancer datasets, and limma and nsFilter feature selection methods in prostate cancer dataset. The classification method with high performance measure values in all three data sets is DL. In general, it may be preferable to use the limma feature selection method in microarray gene expression data with a large number of features. The DL method has given more successful results than classical data mining methods in the classification of large-scale data such as microarray gene expression data and it is recommended to be used.

For future studies, it is planned to use the DL method in different areas such as the recognition of data obtained with medical imaging devices, especially in the field of genetics. Clustering and association rules, which are other frequently used data mining methods, can also be applied on microarray gene expression data.

References

1. Öğüş E. To Be Together Medicine And Biostatistics İn History: Review. Turkiye Klinikleri J Biostat. 9(1):74-83,2017.
2. Öner TÖ, Can Ş. Sağlıkta Biyoistatistiksel Uygulamalar. İzmir Kâtip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi. 3(1):39-45,2018.
3. Karabulut E, Karaağaoğlu E. Biyoinformatik ve Biyoistatistik. Hacettepe Tıp Dergisi. 2010;41:162-170.
4. Polat M, Karahan AG. Multidisipliner Yeni Bir Bilim Dalı: Biyoinformatik ve Tıpta Uygulamaları. S.D.Ü. Tıp Fak. Derg. 16(3):41-50,2009.
5. Yoldaş A, Karaboz İ. DNA Mikroarray Teknolojisi ve Uygulama Alanları. Elektronik Mikrobiyoloji Dergisi TR. 8(1):1-19,2010.
6. Baykara O. Kanser Tedavisinde Güncel Yaklaşımlar. Balıkesir Sağlık Bilimleri Dergisi. 2016;5(3):154-165.
7. Demircioğlu HZ, Bilge HŞ. Yumurtalık Kanseri Veri Kümesindeki Gen İfadelerinin Veri Madenciliği İle Analizi. Marmara Fen Bilimleri Dergisi. 4:125-134,2015.
8. Coşkun E, Karaağaoğlu E. Veri Madenciliği Yöntemleri ile Mikrodizilim Gen İfade Analizi. Hacettepe Tıp Dergisi. 42:180-189,2011.
9. Dolgun MÖ. Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının; Bağımlı Değişken Prevelansı, Örneklem Büyüklüğü ve Bağımsız Değişkenler Arası İlişki Yapısına Göre Karşılaştırılması [Doktora tezi]. Ankara: Hacettepe Üniversitesi; 2014.
10. Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. 3.Baskı. ABD: Elsevier; 2012.
11. Poyraz O. Tıp`Da Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi [Yüksek Lisans Tezi]. Edirne: Trakya Üniversitesi; 2012.
12. Toprak U. Karsinogenezde Mutasyonlar Arası İlişkilerin Veri Madenciliği Metotları İle Tespiti [Yüksek Lisans Tezi]. Trabzon: Karadeniz Teknik Üniversitesi; 2015.
13. Luscombe NM, Greenbaum D, Gerstein M. What İs Bioinformatics? An İntroduction And Overview. Yearbook Of Medical Informatics. 10(01):83-100,2001.
14. Tanır D. Genomik Veri Tabanlarında İndeksleme ve Arama Yöntemleri Üzerine [Doktora Tezi]. İzmir: Ege Üniversitesi; 2017.
15. Sarıkaş A, Odabaşıoğlu N, Altay G. Gen İfade Verilerinde Eksik Değerleri Düzelten Kestirim Yöntemlerinin Karşılaştırılması Comparison of Estimation Methods for Missing Value Imputation of Gene Expression Data. Tıp Teknolojileri Kongresi; 27-29 Ekim; Antalya. s.114-117,2016.
16. Sassanfar S, Walker G. DNA Microarray Technology. What Is It and How Is It Useful, MIT, Biology Science Outreach. 2003.
17. İdil NB. Gen İfade Verileri ile İşlemsel Kanser Sınıflandırılması [Yüksek Lisans Tezi]. Ankara: Başkent Üniversitesi; 2009.
18. Gershon D. Microarray Technology: An Array of Opportunities. Nature. 885-891,2002.
19. Özkan Y, Selçukcan Erol Ç. Biyoenformatik DNA Mikrodizi Veri Madenciliği. 2. Baskı. İstanbul: Papatya Yayıncılık; 2017.
20. Jagota A. Microarray Data Analysis and Visualization. Bioinformatics by the Bay Press: Santa Cruz; 2001.

21. Budak H. Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Süleyman Demirel University Journal of Natural and Applied Sciences. 22 (Special Issue):21-31, 2018.
22. Yazıcı B, Yaslı F, Yıldız Gürleyik H, Turgut UO, Aktas MS, Kalıpsız O. Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama. 9. Ulusal Yazılım Mühendisliği Sempozyumu; 15 - 17 Eylül; İzmir. UYMS-15. s.72-83, 2015.
23. Kaya M. Gen İfade Verilerinde Öznitelik Seçimi ve Sınıflandırma [Yüksek Lisans Tezi].Ankara: Gazi Üniversitesi; 2014.
24. Daş B, Türkoğlu İ. DNA Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması. Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu; 27 – 29 Kasım; Bursa. Eleco, s.381-383,2014.
25. Akman M, Genç Y, Ankaralı H. Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama. Turkiye Klinikleri J Biostat. 3(1):36-48,2011.
26. Çelik N. Amiyotrofik Lateral Skleroz (ALS) Hastalığının Genetik ve Klinik Veri İlişkisinin Veri Madenciliği Yöntemleri ile İncelenmesi [Yüksek Lisans Tezi]. Antalya: Akdeniz Üniversitesi; 2017.
27. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 58(1):267-288,1996.
28. Var E, İnan A. Sınıflandırma İçin Diferansiyel Mahremiyete Dayalı Öznitelik Seçimi. Journal of the Faculty of Engineering and Architecture of Gazi University. 33(1):323-336,2018.
29. Fontı V. Feature Selection using LASSO. Vrije Universiteit Amsterdam; s.26,2017.
30. Ludwig N, Feuerriegel S, Neumann D. Putting Big Data Analytics to Work: Feature Selection for Forecasting Electricity Prices Using the LASSO and Random Forests. Journal Of Decision Systems.;24(1):19-36, 2015.
31. Wibawa AP, Kurniawan AC, Murti DM, Adiperkasa RP, Putra SM, Kurniawan SA ve ark. Nugraha YR. Naive Bayes Classifier for Journal Quartile Classification. iJES. 7(2):91-99,2019.
32. Çataloluk H. Gerçek Tıbbi Veriler Üzerinde Veri Madenciliği Yöntemlerini Kullanarak Hastalık Teşhisi [Yüksek Lisans Tezi]. Bilecik: Bilecik Üniversitesi; 2012.
33. Korkem E. Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2013.
34. Onan A, Korukoğlu S. Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Değerlendirilmesi. XVIII. Akademik Bilişim Konferansı; 30 Ocak-5 Şubat; Adnan Menderes Üniversitesi, Aydın, 2016.
35. Aydın Haklı D. Sınıf Dengesizliği Sorununu Çözmek için Kullanılan Algoritmaların Farklı Sınıflandırma Yöntemlerinde Performanslarının Karşılaştırılması. Ankara: Hacettepe Üniversitesi; 2018.
36. Muthukrishnan R, Rohini R. LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning. IEEE International Conference on Advances in Computer Applications; 2016; Coimbatore. ICACA, s.18-20,2016.
37. Gümüş E. Makina Öğrenme Yöntemleriyle Genom Dizilim Verilerinin Analizi [Doktora Tezi]. İstanbul: İstanbul Üniversitesi; 2013.
38. Fix E, Hodges JL. Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report 4. USAF School of Aviation Medicine, Randolph Field, Texas;1951. Report No:4.
39. Cover MT, Hart P. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. 13(1):21-27, 1967.
40. Elasan S. Veri Madenciliğinde Farklı Karar Ağaçları ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları ve Doğum Verisinde Bir Uygulama [Doktora Tezi]. Van: Van Yüzüncü Yıl Üniversitesi; 2019.
41. Taşcı E, Onan E. K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi. XVIII. Akademik Bilişim Konferansı; 30 Ocak-5 Şubat; Adnan Menderes Üniversitesi, Aydın, 2016.
42. Öğüş E, Can MB, Çamur E, Koru M, Özkan Ö, Rzayeva Z. Veri Kümelerinden Bilgi Keşfi: Veri Madenciliği; 15. Ulusal Biyoistatistik Kongresi, Uluslararası Katılımlı; 20-23 Ağustos; Aydın, 2013.
43. Daş B, Türkoğlu İ. DNA Dizilimlerindeki Nükleotit Çiftlerinin Frekans Değerlerine Göre Farklı Sınıflandırma Yöntemleri ile Karşılaştırılması. Tıp Teknolojileri Ulusal Kongresi; 25 – 27 Eylül; Kapadokya. TıpTekno`14. s.191 -194, 2014.
44. Demircioğlu HZ. Biyoinformatikte Çok Boyutlu Verilerin Boyut İndirgenerek Sınıflandırılması [Yüksek Lisans Tezi]. Ankara: Gazi Üniversitesi; 2015.

45. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science. 313(5786):504-507,2006.
46. Türkçetin AÖ. Akciğer Kanserinin Tespit Edilmesinde Derin Öğrenme Algoritmalarının Kullanılması [Yüksek Lisans Tezi]. Isparta: Isparta Uygulamalı Bilimler Üniversitesi; 2019.
47. Çoşkun C, Baykal A. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. XIII. Akademik Bilişim Konferansı. 2-4 Şubat 2011.
48. Kılıçkap M. Bilgi Kuramı Yaklaşımı ile Bilgisayarlı Tomografik Koroner Anjiyografinin Tanısal Değerinin Değerlendirilmesi [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2012.
49. Dişçi R. Tanı Testlerinin Değerlendirilmesi ROC Analizi. İstanbul Üniversitesi, Onkoloji Enstitüsü; 2013.
50. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology. 2004;3(1):1-26.
51. Çiftçi F, Kaleli C, Günal S. Öznitelik Seçme ve Makine Öğrenmesi Yöntemleriyle Eğitmen Performansının Tahmin Edilmesi. Anadolu Journal of Educational Sciences International. 8(2): 419-440,2018.
52. Smyth GK, Hu Y, Ritchie M, Silver J, Wettenhall J, McCarthy ve ark. limma: Linear Models for Microarray Data [İnternet]. Eylül, 2019. Erişim adresi: https://bioconductor.org/packages/release/bioc/manuals/limma/man/limma.pdf.
53. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. limma: Linear Models for Microarray and RNA-Seq Data User's Guide [İnternet]. Kasım, 2019. Erişimadresi:https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf.
54. Pala T. Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi [Yüksek Lisans Tezi]. İstanbul: Marmara Üniversitesi; 2013.
55. Vapnik VN. An Overview of Statistical Learning Theory. IEEE Transactıons on Neural Networks.1999;10(5):988-999.
56. Kaşıkçı M. Transkriptom Veri Seti Üzerinde Derin Öğrenme Yöntemi ile Klasik Veri Madenciliği Yöntemlerinin Sınıflama Performanslarının Karşılaştırılması [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2019.
57. Kayaalp F, Başarslan MS, Polat K. Kronik Böbrek Hastalığını Tanımlamada Bir Hibrit Sınıflandırma Örneği**.** Electric Electronics, Computer Science, Biomedical Engineerings Meeting;18-19 Nisan 2018; İstanbul Arel Üniversitesi, İstanbul.
58. Doğan F, Türkoğlu İ. Derin Öğrenme Algoritmalarının Yaprak Sınıflandırma Başarımlarının Karşılaştırılması. Sakarya University Journal of Computer and Information Sciences.2018;1:10-21.
59. Arik Ö. (2020), Mikrodizi Gen İfade Verilerinde Farklı Öznitelik Seçim Yöntemleri İle Sınıflama Yöntemlerinin Performanslarının Değerlendirilmesi. Doktora Tezi, Hacettepe Üniversitesi, Ankara

S19

Performance Investigation Of Approaches And Classification Methods Used In Integration Multi-Omics Data

Funda İpekten[1,2], Gözde Ertürk Zararsız[1,2], Halef Okan Doğan[3], Vahap Eldem[4], Gökmen Zararsız[1,2]

[1]Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey
[2]Drug Application and Research Center, Erciyes University, Kayseri, Turkey
[3]Department of Biocmehistry, Faculty of Medicine, Cumhuriyet University, Sivas, Turkey
[4] Department of Biology, Faculty of Sciences, Istanbul University, Istanbul, Turkey

fundaipekten@gmail.com

Abstract

Introduction: The use of multi-omic technologies allows one to examine systems biology, cellular metabolism, and disease etiology in more detail. More powerful and easily interpretable results can be obtained using multi-omic technologies instead of a single omic technology to analyze the molecular complexity of the disease`s etiology. Therefore, it is thought that the combining and classification methods used in omic data can provide significant contributions in terms of models that can be used in the biological system and diagnosis and treatments of the disease. This study aims to evaluate the performance of data integration and classification methods on multi-omic data.

Method: In the study, three different real data sets, namely colon, kidney, and thyroid cancer, were used. The raw data generated using mRMR and PCA analysis were combined using concatenate integration, transformation-based integration, and model-based integration methods. The classification methods of CANetwork estimated the subtype of the disease. The subtype of the disease were estimated by the classification methods of CANetwork, RVM, and Ada-boost RVM in the data combined with the MKL, NSC, RF, and SVM classification methods data combined with the transformation-based integration method.

Results: The classification performance of the RVM method was found to be the highest in the transformation-based integration kidney data (AUC=0.927). The method with the highest classification performance in colon, kidney, and thyroid data obtained by applying mRMR variable selection in concatenation data is MCL (AUC=1,000). In the column data obtained using mRMR variable selection in model-based combined data, the method with the highest classification performance is RF (AUC=1.000), while SVM in kidney and thyroid data (AUC values of the data are 0.967, 0.669, respectively). In the kidney data obtained by applying the mRMR variable selection and PCA method in the concatenation data, the method with the highest classification performance was MKL (AUC=0.926), while the SVM in the colon and thyroid data (AUC values of the data were 0.603, 0.582, respectively). SVM has the highest classification performance in colon, kidney, and thyroid data obtained by applying mRMR variable selection and PCA method in model-based integrated data (AUC values of the data are 0.809, 0.950, 0.661, respectively).

Conclusion: The application results obtained from the real data sets may differ in the superiority of the performances of the data and methods used compared to each other. Combining the data using mRMR variable selection with the concatenation method has been shown that the MCL classification method of these data can separate the subtypes of the diseases more accurately than other methods.

Keywords: Multi-omic, metabolomics, RNA-sequencing, next-generation sequencing

S20

Comparison of the Effects of Median and Cyclic Loess Normalization Methods Used in Microarray Data on Deep Learning Performance

Asena Ayça ÖZDEMİR[1], Gülhan TEMEL[2], Saim YOLOĞLU[3]

[1] *Mersin University Department of Medical Education, Mersin,*
[2] *Mersin University Department of Biostatistics and Medical Information, Mersin,*
[3] *Inonu University Department of Biostatistics and Medical Informatics, Malatya*

Introduction: Many different normalization methods are applied to help reduce measurement error in microarray data. Median and Cyclic Loess methods are used to normalize microarray data. In addition, especially for classification purposes, deep learning method has been used quite frequently for gene expression data like structures of many data in recent years. Before applying the deep learning method, normalization is recommended to prepare data of gene expression for analysis. The aim of this study was to compare the deep learning classification performance results of Median and Cyclic Loess normalization methods used in single color microarray gene expression data.

Method: 1.000 single-color microarray data were generated from each data group with sample widths of 5:5, 20:20, 100:100 and 100, 1.000, 10.000 genes as two independent groups in the R Madsim package. Median and Cyclic Loess normalization methods were applied to the generated data in R. Deep learning algorithms were applied to raw data and normalized datasets in Python. The results were interpreted by taking the average of classification performances of 1000 datasets.

Results: The mean accuracy obtained via deep learning from 100 genes were ranked from high to low, Cyclic Loess (0.689±0.254), Median (0.688±0.258), raw data (0.681±0.252) in 5:5 dataset; raw data (0.718±0.176), Cyclic Loess (0.710±0.177), Median (0.709±0.178) in 20:20 dataset; Median (0.731±0.183), raw data (0.728±0.186), Cyclic Loess were calculated as (0.725±0.180) in the 100:100 dataset. The mean accuracy obtained via deep learning from 1.000 genes were ranked from high to low; raw data (0.848±0.193), Median (0.843±0.185), Cyclic Loess (0.840±0.186) were found for 5:5 dataset, raw data (0.850±0.153), Cyclic Loess (0.849±0.148), Median (0.848±0.155) for 20:20 dataset and Cyclic Loess (0.879±0.178), raw data (0.875±0.182), and Median (0.874±0.184) were calculated for 100:100 dataset. The mean accuracy obtained from deep learning for 10.000 genes were also ranked from high to low; raw data (0.628±0.247), Median (0.627±0.244), Cyclic Loess (0.614±0.243) for 5:5 dataset; raw data (0.597±0.197), Cyclic Loess (0.591±0.194), Median (0.588±0.190) for 20:20 dataset and Cyclic Loess (0.596±0.194), Median (0.594±0.192), raw data (0.594±0.198) were calculated for 100:100 dataset. For each sample and gene size, no statistically significant result was obtained in terms of the mean accuracy obtained from raw data of deep learning and normalization methods.

Conclusion: It has been observed that the normalization methods applied in single-color microarray datasets have no effect on deep learning performance. Sample size, also, have no significant effect on deep learning performance, although deep learning performance increases when size of the gene is not very small (100 genes) or large (10.000 genes).

Keywords: Deep Learning, Microarray, Normalization

S21

Copula-Based Gene Selection for Survival Prediction

Aslıhan Alhan[1]

[1] *Ufuk University Faculty of Medicine, Department of Biostatistics, Ankara*

Introduction: Genetic information is noteworthy in the prediction of survival. Gene selection based on the Cox regression model traditionally applied in research to select genes that predict survival has based on the assumption of independent censoring. However, in the case where censorship is dependent, there may be bias in gene selection, and there may be deviation in the efficiency of the model obtained. This study aims to introduce the Copula-based gene selection method, which offers an alternative approach to gene selection, taking into account the effect of the situation on which censoring is dependent. In addition, this method is suggested as an alternative approach for gene selection and personalized patient survival prediction.

Method: Chen et al.'s (2007) survival data of 125 non-small cell lung cancer patients were analyzed using the "compound.Cox" R package program.

Results: 38 patients died, and the remained 87 patients have censored during the follow-up period of approximately five years in the non-small cell lung cancer data. Moreover, the 125 patients were divided into two groups as 63 training and 62 patients for testing to evaluate by the cross-validation method of the performed a prognostic index and genes selection of the data of 63 patients in the training group. The prognostic index- a value that can evaluate by routine examination and examination findings for ease of application- was calculated with the selected genes. Clayton copula and Copula-graphic models were employed to find the predictive performance of the prognostic index.

Conclusions: Taking into explanation the effect of Clayton copula-dependent censoring, 16 genes (MMP16, ZNF264, HGF, HCK, NF1, ERBB3, NR2F6, AXL, CDC23, DLG2, IGF2, RBBP6, COX11, DUSP6, ENG, IHPK1) selected among 97 genes in the prognostic index model. In addition, HCK, IGF2, ENG, and IHPK1 genes were found as protective genes while others as risk genes. Two survival curves were assessed by adjusting the effect of dependent variables with the Clayton copula model and Copula-graphic model to find the predictive performance of the prognostic index. As a result of this analysis, the measurement between good and bad prognoses was significant (mean difference = 0.224; p = 0.021). This result appears to confirm the predictive ability of the prognostic index obtained using the copula-based gene selection approach.

Keywords: Survival prediction, gene selection, dependent censoring, copula.

References:

1. Emura, T ve Chen, YH. (2018), Analysis of Survival Data with Dependent Censoring Copula-Based Approaches. Springer.
2. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY et al (2007) A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 356:11–20
3. Emura T, Chen YH (2016) Gene selection for survival data under dependent censoring, a copula-based approach. Stat Methods Med Res 25(6):2840–2857
4. Emura T, Nakatochi M, Matsui S, Michimae H, Rondeau V (2017) Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model. Stat Methods Med Res, https://doi.org/10.1177/0962280216688032
5. Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York
6. Rivest LP, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. J Multivar Anal 79:138–155.

S22

Sağkalım Analizinde Zamana Bağlı Açıklayıcı Değişkenler için Birleşik Model Yöntemi Kullanımı

(Using the Joint Model Method for Time-Related Response Variables in Survival Analysis)

[1]**Fatma KAYMAKAMTORUNLARI DENİZ,** [2]Osman Tolga KASKATI, [3]Merve BAŞOL GÖKSÜLÜK

[1]Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara.
[2] Lokman Hekim Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara.
[3]Erciyes Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Kayseri.

Özet

Giriş: Sağkalım analizi, ilgilenilen bir olay ile karşılaşıncaya kadar geçen zamanı analiz eden ve bu zaman üzerinde etkisi olan değişkenlerin belirlenmesine ve incelenmesine olanak sağlayan istatistiksel bir yöntemdir. Klinik denemelerde ya da izlem çalışmalarında oldukça yaygın olarak kullanılan sağkalım analizlerinde Cox regresyon modeli sıkça kullanılmaktadır. Cox regresyon modeli (veya oransal hazard regresyonu), ilgilenilen olayın meydana geldiği süre üzerinde etkili olabilecek değişkenleri araştırmak için kullanılan bir yöntemdir. İlgilenilen olay, ölüm, hastalık gelişimi, nüks etme veya diyalize girme gibi durumlar olabilir. Genel Cox regresyon modellerinde başlangıç risk fonksiyonu zamanın bir fonksiyonunu içerirken, açıklayıcı değişkenler genellikle zamandan bağımsız olarak alınır. Ancak, izlem süresi içerisinde kişilerin hastalık durumlarını (ilerleme ya da gerileme) takip edebilmek adına bu kişilerden tekrarlı olarak ölçümler alınabilir. Bu çalışmada, Cox regresyon modelinde zamandan bağımsız ve zamana bağlı açıklayıcı değişkenlerin olduğu durumlarda kullanılan birleşik model yaklaşımı tanıtılmış ve R programlama dili kullanılarak gerçek bir veri seti üzerinde uygulama yapılmıştır. Bu sayede birleşik model yaklaşımının kullanımının arttırılması hedeflenmiştir.

Yöntem: Tekrarlı ölçüm verileri, uzunlamasına veriler olarak adlandırılır. Uzunlamasına verilerde, aynı kişiden birden çok ölçüm alındığı için ölçüm değerleri arasında bağımlılık durumu söz konusudur. Bu bağımlılık durumunu da dikkate alarak sağkalım verileri ile uzunlamasına verilerin birlikte analiz edilmesinde birleşik model yöntemi kullanılmaktadır. Birleşik model iki alt modelden oluşur: (i) uzunlamasına veri modeli ve (ii) sağkalım modeli. Birleşik model kullanılmasıyla izlem süresi içerisinde tekrarlı olarak alınan tüm ölçümlerin modele katkısı dikkate alınarak modelleme yapılması sağlanır. Bu çalışmada, Ankara Üniversitesi Tıp Fakültesi Nefroloji Bilim Dalı'ndan 2015-2021 yılları arasında retrospektif olarak elde edilen, nefrotik sendromlu 88 hastanın demografik ve klinik verilerinin kaydedildiği veri seti kullanılmıştır. Çalışmada, diyaliz durumunu ile proteinüri değerleri arasında ilişki olup olmadığı incelenmiştir. Proteinüri değerleri 0.ay, 3.ay, 6.ay, 12.ay ve 24.ayda alınmıştır. İlk aşamada proteinüri değerleri doğrusal karışık etkili model kullanılarak modellenmiş ve bu modelden elde edilen kestirim değerleri sağkalım modeline dahil edilmiştir.

Bulgular ve Sonuç: Nefrotik sendromlu 88 hastanın 18 (%20.5)'nde diyaliz olayı görülürken, 70 (%79.5)'nde diyaliz olayı yoktu. Diyalize giren hastaların yaş ortalaması 54.78±14.20 iken diyalize girmeyenlerin yaş ortalaması 54.63±12.46 idi. Yaş ve VKİ bakımından gruplar benzerdi. Diyalizde olan 18 hastanın %83.3'ü erkek, %16.7'si ise kadındı. Proteinüri değerlerini etkileyen açıklayıcı değişkenleri belirlemede tek değişkenli doğrusal karışık etkili model kullanılmış, bu modelden elde edilen değişkenler çoklu doğrusal karışık etkili modele alınmıştır. Protenüri'yi etkileyen değişken olarak sadece remisyon durumu anlamlı bulunmuştur (p=0.002). Renal Sağkalıma etki eden risk faktörlerini belirlemek için ilk olarak tek değişkenli Cox regresyon modeli yapılmış sonrasında ise çok değişkenli Cox orantılı risk modeli kullanılmıştır. Çok değişkenli Cox analizi sonucunda diyalize etki eden değişken, geliş anında renal replasman tedavi ihtiyacı (yatışta RRT) olup olmama durumu belirlendi. Birleşik modelde ise doğrusal karışık etkili model sonucunda elde edilen proteinüri kestirim değerleri sağkalım alt modeline eklenerek birleşik model uygulanmıştır. Birleşik model sonucunda proteinüri değerlerinin renal sağkalım üzerinde anlamlı bir etkisi saptanmazken, yatışta RRT modelde anlamlı olarak saptanmıştır (RR=10.64[3.22-32.99], p<0.001). Birleşik model kullanılmasıyla izlem süresi içerisinde tekrarlı olarak alınan tüm ölçümlerin modele katkısı dikkate alınarak modelleme yapılması ve kayıp verilerden etkilenmiyor olması nedeniyle de bu tür verilerin analizinde birleşik model kullanılmasının daha bilgi verici sonuçlar elde edilmesine olanak sağladığı görülmüştür.

Anahtar Sözcükler: Cox regresyon, uzunlamasına veriler, birleşik model

Kaynaklar

1. Alsefri M, Sudell M, García-Fiñana M, Kolamunnage-Dona R. Bayesian joint modelling of longitudinal and time to event data: a methodological review. BMC Medical Research Methodology. 2020;20(1):94.
2. Elashoff RM, Li G, Li N. Joint Modelling of Longitudinal and Time-to-Event Data. Chapman and Hall/CRC; 2016.
3. Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data with Applications in R. Chapman and Hall/CRC; 2012.
4. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics. 2011 Sep;67:819–829.
5. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2010 Jun;28:2796–2801.
6. Proust-Lima C, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics (Oxford, England). 2009 Jul;10:535–549.
7. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. Statistics in Medicine. 1996;15(15):1663–1685.

S23

Data Format for Rapid Biostatistics Analysis: PARQUET

SU ÖZGÜR[1,2]

[1]*EgeSAM-EgeTPRC (Ege University Translational Pulmonary Research Center) Bornova, İzmir/TURKEY*

[2]*Regional Hub for Cancer Registration in Northern Africa, Central and Western Asia, WHO/IARC- GICR*

SUMMARY

Introduction: Very large data stacks obtained from different sources in the health field have an effective potential for issues such as defining the requirements between the patient and user/systems that provide health care services and optimizing the services provided. Big data in healthcare is defined as the tools, technologies, and processes used to appropriately create, store, process, and analyze electronic medical data (EMR) **(1,2)**. For this reason, institutions and organizations who are providing health services; they can use big data as a baseline tool to improve the quality of health services, including identifying deficiencies in the system, identifying risks, reducing human and non-human-induced medical errors, control/improve the accuracy of the health outputs produced, reducing the costs to the health system and increasing employee/patient satisfaction. Technological developments enable the development of new data analysis methods and systems that will support clinical decision-making. These developments will enable new care and treatment methods beyond the traditional patient experience by processing clinical, genomic, imaging, biometric, and epidemiological data accurately and efficiently **(3)**. Storage media and stored data format are important for rapid evaluation of data. Parquet is a columnar storage format, designed to support fast querying, very efficient compression and encryption of schemas compared to row-based files such as CSV.

Aim: In this study, the advantages and disadvantages of the usage of parquet format were investigated in order to rapidly merge and query different data formats (genetic and clinical) in the field of health. It is aimed to show how much the analysis speed and reliability of the "1000 Genomes Project" data, which reaches millions of column sizes, have changed with simulated clinical data in parquet format.

Method: The genetic data of 2504 individuals from 26 populations representing the European, East Asian, South Asian, West African and American populations in the main 1000 Genomes Project **(4)** were converted to parquet format **(5)**. Demographic dataset in the same project was added to the converted data in parquet format. SQL queries were defined for the final data table.

Results: In the study, number of query comparison of classical raw genetic and clinical data with data in parquet format was converted. Especially for cohort studies, it was calculated that the speed of creating sub-data tables increased by an average of 70%. This data format can be integrated into decision support systems when real-time analyses are required in population-based studies. Furthermore, it was found that the model (machine learning, etc.) accelerated the training process by 28% on average.

Conclusion: The use of non-traditional data formats should be considered to satisfy different needs in healthcare, clinic, and medical research. The Parquet data format has a simple structure, faster writing efficiency to databases. Parquet is most compatible with machine learning processes too. Therefore, for large-scale analytical projects, converting data to Parquet format is important in terms of cost and rapid findings. Considering the costs of using cloud computing environments, this increase in efficiency will save time, effort and money of researchers.

***Keywords**: Parquet, EMR, Cloud computing*

References

1. Olaranke I, Oluwaseun O. Big Data in Healthcare: Prospects, Challenges and Resolutions. FTC 2016 - Future Technologies Conference'2016. 6-7 December 2016 . San Francisco, United States.
2. *Büyük Verinin Sağlık Hizmetleri Kalitesindeki Rolü*. Available from: https://www.researchgate.net/publication/326553653_Buyuk_Verinin_Saglik_Hizmetleri_Kalitesindeki_Rolu [accessed Sep 29 2021].

3. Stanford Medicine (2017). Health Trends Report: Harnessing the Power of Data in Health, July 2017. https://med.stanford.edu/content/dam/sm/smnews/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf [accessed Sep 30 2021].
4. Sudmant, P., Rausch, T., Gardner, E. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015). https://doi.org/10.1038/nature15394
5. https://github.com/microsoft/genomicsnotebook/tree/main/vcf2parquet-conversion

S24

A Comprehensive Simulation Study To Determine The Most Appropriate Regression And Confidence Interval Method İn Method Comparison Studies

Gözde Ertürk Zararsız[1,2], Gökmen Zararsız[1,2], Dinçer Göksülük[1], Cengiz Bal[3], Halef Okan Doğan[4], Ahmet Öztürk[1], Gabi Kastenmüller[5]

[1]*Erciyes University, Faculty of Medicine, Department of Biostatistics, Kayseri, Turkey*
[2]*Erciyes University, Pharmaceutical Research, and Application Center (ERFARMA), Kayseri, Turkey*
[3]*Eskişehir Osmangazi University, Faculty of Medicine, Department of Biostatistics, Eskişehir, Turkey*
[4]*Sivas Cumhuriyet University, Faculty of Medicine, Department of Biochemistry, Sivas, Turkey*
[5]*Helmholtz-Zentrum München, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany*
gozdeerturk@erciyes.edu.tr

Introduction: Method comparison studies are frequently used by manufacturers and laboratory specialists to determine the measurement error by comparing the reference and test methods of in vitro diagnostic measurements. This study evaluates regression methods performance and confidence interval approaches in different scenarios by performing a comprehensive simulation for cases where measurement errors have constant variance. In particular, it aimed to investigate the performance of different bootstrap confidence interval approaches in method comparison studies and obtain guiding findings for researchers.

Method: In the study, least squares (LC), Deming (DR), and Passing-Bablok (PB) regression methods and analytical, jackknife, bootstrap percentile, bootstrap student, bootstrap Bca, and bootstrap t confidence interval approaches were used. Based on the findings from real data, a comprehensive simulation was organized. The performances of the regression methods and confidence interval approaches were investigated and compared with each other in simulation schemes containing all possible combinations of the different distribution range, sample size, measurement distribution, analytical standard deviation rate, whether the measurement error rate is known, whether there is a practical observation or not, the presence of constant and proportional error. In these comparisons, type-I error rates and strengths of the methods were determined as evaluation criteria. For this purpose, each simulation was repeated 5,000 times. The bootstrap number was selected as 999.

Results: The performances of the DR and PB methods were higher than the EKK method. It has been observed that the DR method gives the best results when the measurement error is known, the analytical standard deviation ratio is 1, the distribution range is comprehensive, and there are no influential observations. The best results were obtained when the analytical standard deviation ratio for the PB method was 1. Furthermore, when jackknife, bootstrap percentile, bootstrap Bca, and bootstrap t confidence intervals were used for the DR method, the methods had the best performance when bootstrap percentile and bootstrap Bca confidence intervals were used for the PB method.

Conclusion: In method comparison studies, the ECC method should not be preferred in cases where both methods have measurement errors. Although DR and PB methods have the best performance; More care should be taken in cases where the measurement error is unknown, the distribution range is narrow, the analytical standard deviation ratio is not 1, and there is effective observation. The best results can be achieved using jackknife and bootstrap confidence intervals with the DR method and the bootstrap confidence intervals with the PB method. It is recommended to use bootstrap confidence intervals, especially in cases where the number of observations is low.

Keywords: Bootstrap, linear regression, in vitro diagnosis, systematic error, method comparison

S25

Box-Cox Dönüşüm Parametresi Kestirimi İçin Yeni Bir Yaklaşım

Muhammed Ali YILMAZ[1], Osman DAĞ[2]

[1] Sağlık Bakanlığı Çağrı Merkezi ve İletişim Daire Başkanlığı, Ankara
[2] Hacettepe Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara

Giriş: Çoğu istatistiksel yöntem normal dağılım varsayımı altında geliştirilmiştir. Fakat gerçek hayatta veriler genellikle normallik varsayımını sağlamayabilir. Bu koşulun sağlanabilmesi için veri üzerinde farklı dönüşüm yöntemleri uygulanabilmektedir ve bu yöntemler arasında en çok kullanılan yöntemlerden biri ise Box-Cox dönüşüm yöntemidir. Bu çalışmada farklı yöntemlerle kestirilen Box-Cox dönüşüm parametrelerini meta analiz yoluyla bir araya getiren yeni bir algoritma önerilmiştir. Daha sonra Monte-Carlo benzetim çalışması ile farklı senaryolar altında önerilen algoritmanın diğer yöntemlere göre performansı değerlendirilecektir.

Amaç: Dönüşüm yöntemleri arasında önemli bir yeri olan Box-Cox yöntemini uygularken karşılaşılan önemli problem dönüşüm parametresinin kestirimidir. Literatürde bilinmeyen dönüşüm parametresinin kestirimi için birçok çalışma yapılmıştır. Bu çalışmada ise performans olarak daha iyi bir kestim yöntemi geliştirmek amaçlanmıştır.

Yöntem: Parametre kestirimi için öncelikle bir lambda aralığı belirlenmelidir. Daha sonra Asar vd. (2017)'nin çalışmasında önerilen algoritma kullanılarak Shapiro-Wilk, Anderson–Darling and Jarque–Bera testleriyle lambda parametreleri kestirilir. Parametrik olmayan bootstrap yöntemiyle ise kestirimlerimizin standart hataları elde edilir. Daha sonra meta analizinde kullanılan rastgele etki modelini yardımıyla ağırlıklandırılmış ortalamayla parametre kestirimi yapılır. Burada ağırlıklandırma standart hata ile ters orantılıdır. Son olarak da farklı normallik testleriyle elde edilen kestirimlerin meta analizindeki rastgele etki modeliyle ağırlıklandırılmış ortalaması elde edilir.

Bulgular: Önerilen yöntemin etkinliği ile literatürde önerilen yöntemleri karşılaştırabilmek için yanlılık, standart hata ve hata kareler ortalamasından yararlanılmıştır. Genel olarak bakıldığında yanlılık ve hata kareler ortalaması, dönüşüm parametresinin değeri arttıkça daha da azalmaktadır. Fakat örneklem büyüklükleri azaldıkça diğer kestirim yöntemleri arasındaki fark daha da netleşmektedir.

Sonuç: Sonuç olarak bu çalışmada Box-Cox yönteminde bilinmeyen dönüşüm parametresinin kestirimi için yeni bir yöntem önerilmiştir. Meta analizindeki rastgele etkiler modeli ile Asar vd. (2017)'nin yaptığı çalışmadaki en iyi performansı gösteren normallik testlerinin altı tanesinin farklı kombinasyonlarında parametre kestirimlerini birleştirdik. Box-Cox parametresi kestirimi için sadece bu yöntemlerle elde edilen kestirimlerin ağırlıklandırılmış ortalamaları yerine mümkün tüm kombinasyonlarından (toplamda 63 farklı kombinasyon) elde edilen tahminlerin ağırlıklandırılmış ortalamaları alındı. Yapılan Monte-Carlo benzetim çalışması sonucunda da en iyi performans gösteren kombinasyonun, Shapiro-Wilk, Anderson–Darling and Jarque–Bera testleriyle elde edilen kombinasyon olduğu saptanmıştır. Box-Cox dönüşümü için parametre kestirim yöntemleriyle karşılaştırıldığında yeni önerdiğimiz yöntemin diğer yöntemlere göre daha etkili sonuçlar verdiği görülmektedir.

Anahtar Sözcükler: Box-Cox dönüşümü, Monte Carlo benzetim çalışması, normality tests.

Kaynaklar

Asar, O., Ilk, O., Dag, O. (2017). Estimating Box-Cox power transformation parameter via goodness-of-fit tests. Communications in Statistics-Simulation and Computation, 46(1), 91-105.

S26

Meta Analizinde Yanlılığı Düzeltme Yaklaşımları

Emre Dirican[1]

[1]Hatay Mustafa Kemal Üniversitesi, Tayfur Ata Sökmen Tıp Fakültesi, Biyoistatistik Anabilimdalı

Alahan, Tayfur Sökmen Kampüsü, 31060 Antakya/Hatay

Giriş: Yayın yanlılığı, sistematik derleme ve meta analizlerinde sonuçların hem geçerliliğini hem de genellenebilirliğini etkileyebilecek önemli bir sorundur. Çalışmanın amacı yayın yanlılığının olduğu veya toplam etki için kritik sonuçların alındığı durumlarda, bu problemlere çözüm sağlayabilecek yöntemlerin performanslarını kıyaslamaktır.

Yöntem: Çalışmada kırp doldur (trim fill), Copas seçim modeli ve regresyon tabanlı yanlılık düzenleme metotlarına yer verilmiştir. Bu yöntemler, değerlendirmeleri odds ratio, risk ratio, hazard ratio, tek bir ortalama ve standartlaştırılmış ortalamalar fark olan meta analizler üzerinde çalışılmıştır. Yöntemlerin performansı hata kareler ortalaması ve Egger'in yanlılık değerleri bakımından kıyaslanmıştır.

Bulgular: Minimum 7, maksimum 37 çalışmadan oluşan 16 meta analizi ile yanlılık düzenleme yöntemlerinin performansı kıyaslanmıştır. İkili sonuçların değerlendirildiği (odds ratio, hazard ratio, risk ratio ve n=10) meta analizlerinin çoğunda (%70) trim fill metodundan performanslı sonuçlar alınmakla birlikte bazı çalışmalarda (%30) Copas'ın seçim modelinin de performansı dikkat çekmiştir. Regresyon tabanlı yöntemden ise Copas sonuçlarına yakın ancak daha az performanslı sonuçlar alınmıştır. Tek bir ortalama ve standartlaştırılmış ortalamalar arasındaki farkı değerlendiren meta analizlerinde (n=6) trim fill metodu en yüksek performansı göstermiştir.

Sonuç: Çalışmada kullanılan yöntemlerin içerisinden trim fill metodunun, meta analizinde yayın yanlılığı olduğu durumda öncelikle kullanılması gereken yöntem olduğu anlaşılmaktadır. Ancak özellikle ikili sonuçların değerlendirildiği ve kritik sonuçların alındığı meta analizlerinde Copas'ın seçim modelinden elde edilecek sonuçların da değerlendirilmesi gerektiği düşünülmektedir.

Anahtar Sözcükler: Meta analizi, yanlılık, trim fill, Copas

S27

Comparison of Stereotype Logistic Regression and Proportional Odds Logistic Regression Models for Ordered Response Variables: An Application on Hyperopia Data

Hulya OZEN

Eskisehir Osmangazi University Faculty of Medicine, Department of Biostatistics, Eskisehir

Aim: In this study, it is aimed to compare the performances of Proportional Odds Logistic Regression (PO) and Stereotype Logistic Regression (SL) models on the hyperopia dataset containing ordered response variable and to introduce the unique features of the SL model.

Method: The SL model is a flexible approach to modeling ordered response variables. It is a method that does not require the proportional odds assumption, preserves the hierarchy between the response variables with the constraints in its algorithm, and provides a different interpretation of odds ratios. The data set used in this study contains a response variable that includes four levels of hyperopia, and some of the factors that may be associated with the occurrence of these levels such as age, eye pressure, choroidal thickness and largest vessel lumen diameter. PO and SL models were applied on the dataset. The proportional odds assumption was investigated with the Brant test. AIC and BIC evaluation criteria and Cox-Snell R^2 values were used to compare the regression models.

Results: As a result of Brant test, it was observed that the proportional odds assumption was violated ($P<0.001$). As the evaluation criteria are examined; the values of the SL model (AIC:232.21; BIC:257.37) were found to be lower than those of the PO model (AIC: 239.97; BIC: 259.54). Cox-Snell R^2 values were observed as 0.571 for SL and 0.527 for PO. It was obtained that the SL model fitted the data set better. Choroidal thickness ($P<0.001$) and largest vessel lumen diameter ($P=0.002$) were found to significantly affect hyperopia levels according to the PO model. In the SL model, significant variables were determined as choroidal thickness ($P<0.001$), largest vessel lumen diameter ($P=0.004$) and age ($P=0.040$). While the cumulative odds ratios between the levels were calculated with the PO model, the odds ratios of the levels were obtained separately for a certain reference category with the SL model.

Conclusion: This study reveals that SL model can be used as an alternative method in healthcare applications where ordered response variables are frequently used. It provides creating a more flexible model without being affected by the proportional odds assumption.

Keywords: Proportional odds logistic regression, Stereotype logistic regression, Ordinal response variable, Hyperopia

References

1. Agresti, A. (2010).Analysis of ordinal categorical data(2nd ed.)., Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley
2. Anderson, J. A. (1984). Regressionandorderedcategorical variables. Journal of the Royal Statistical Society Series B, 46(1), 1–30
3. Liu, X. (2014). Fitting Stereotype Logistic Regression Models for Ordinal Response Variables in Educational Research (Stata), - Journal of Modern Applied Statistical Methods: V ol. 13 : Iss. 2 , Article 31.
4. Fernandez D, Liu I, Costilla R. (2019). A method for ordinal outcomes: The ordered stereotype model. Int J Methods Psychiatr Res. 28:e1801. https://doi.org/10.1002/mpr.1801
5. Abreu MNS, Siqueira AL, Cardoso CS, Caiaffa WT. (2008) Ordinal logistic regression models: application in quality of life studies. Cad. Saúde Pública. 24(4):581-591
6. Brant R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 1171-1178

S28

Lojistik Regresyon Analizi ile Elde Edilen Beta Katsayısı, Odds Oranı ve Tamsayılı Skor Sisteminin Klinik Tahmin Modeli Geliştirilmesindeki Başarısının Karşılaştırılması

Comparison of the Success of Beta Coefficient, Odds Ratio and Point Score System Obtained by Logistic Regression Analysis in Developing a Clinical Prediction Model

Gülçin AYDOĞDU[1], Emre DEMİR[1], Yasemin YAVUZ[2]

[1] *Hitit Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Çorum*
[2] *Ankara Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara*

Özet

Amaç: Klinik tahmin modelleri yaş, cinsiyet ve biyobelirteçler gibi tahmin edici sayısının birden fazla olduğu durumlarda prognostik veya risk faktörlerini kullanarak bir birey için gelecekteki bir olayın veya hastalığın riskini tahmin etmek amacıyla yaygın olarak kullanılmaktadır. Literatürde yaygın kullanılan birçok önemli skor bulunmaktadır ve her yıl tıp literatüründe çok sayıda tahmin modeli geliştirilmektedir. Bunlara yoğun bakım ünitesinde ölüm riskini değerlendirmek için kullanılan APACHE skorları, Akut apandisit tanısında kullanılan Alvarado skoru, Pulmoner emboli için Wells Kriterleri, 10 yıl içerisinde bireylerin koroner kalp hastalığı riskini öngörmede kullanılan Framingham koroner kalp hastalığı risk skoru ve hastalarda bilinç (nörolojik) durumunun değerlendirilmesinde kullanılan Glasgow Koma Skoru örnek olarak gösterilebilir. Klinik tahmin modellerinin geliştirilmesinde sonuç değişkeninin yapısına bağlı olarak doğrusal, lojistik ve Cox modellerini içeren regresyon analizleri yaygın olarak kullanılmaktadır. Bu çalışmada regresyon katsayılarına ve risk oranlarına dayalı klinik tahmin modellerinin performanslarının karşılaştırması amaçlanmıştır.

Yöntem: Bu çalışmada iki düzeyli sonuç değişkene sahip tahmin modellerinde risk faktörlerinin skor değerlerini belirlemek için modelde kategorik yapıdaki bağımsız değişkenlerin bulunduğu n=1000 örneklem büyüklüğünde üretilen benzetim veri seti ile lojistik regresyon analizi gerçekleştirildi. Tahmin modellerinin pratikte kullanımının kolay olması amacıyla modelde anlamlı bulunan tahmin edici parametreler için lojistik regresyon ile elde edilen β katsayıları ve odds oranları genellikle tamsayılı skorlara dönüştürülmektedir. Çalışmamızda da lojistik regresyon analizinde anlamlı bulunan β katsayıları, odds oranları ve bu değerlerin tamsayıya dönüştürülmesi ile elde edilen ağırlıklı bir skorlama sistemi oluşturulduktan sonra bu skorlara dayalı 4 farklı klinik tahmin modelinin başarıları ROC, eğri altında kalan alan (AUC), duyarlılık, seçicilik, pozitif-negatif tahmin değerleri ve doğruluk değerleri ile karşılaştırıldı. Tüm istatistiksel analizler ve benzetim için R studio (Versiyon 3.5.0) paketi kullanıldı.

Bulgular: Çalışmamızda 4 farklı senaryo için 10000 benzetim sonucunda elde edilen klinik tahmin modellerinin sonuç değişkenini tahmindeki AUC (%95 güven aralığı) performansları β katsayıları, tamsayıya dönüştürülmüş β katsayıları, tamsayıya dönüştürülmüş odds oranları ve odds oranları için sırasıyla 0.768 (0.740-0.797), 0.754 (0.724-0.783), 0.674 (0.641-0.707), 0.670 (0.637-0.703) bulunmuştur.

Sonuç: Klinik tahmin modellerinde henüz pratikte uygulama birliği bulunmamaktadır. Literatürde klinik tahmin modellerinin oluşturulmasında yaygın olarak β katsayıları, odds oranları ve bunların tamsayıya yuvarlanmış skorları kullanılmaktadır. Skorları belirlemek için kullanılan bu katsayılar model tahmin başarısını doğrudan etkilemektedir. Benzetim sonuçlarına göre sadece kategorik değişkenlerden oluşan modellerde ROC analizi sonucunda AUC'ye göre en başarılı klinik tahmin modeli β katsayıları ile oluşturulan model olmuştur. Daha sonra ise AUC'ye göre başarı sırası tamsayıya dönüştürülmüş β katsayıları, tamsayıya dönüştürülmüş odds oranları ve odds oranları olarak belirlenmiştir.

Anahtar Sözcükler: Klinik tahmin modeli, Lojistik regresyon, Odds oranı, tamsayılı skorlama sistem

S29

Seyrek (Sparse) Lojistik Regresyon ve Çok boyutlu Genomik Veri Seti Üzerine Bir Uygulama

Özlem KAYMAZ[1], Khaled ALQAHTANİ[2] ,Henry WOOD[3] , Arief GUSNANTO[4]

[1]Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü, Ankara,Türkiye
[2] Prince Sattam Bin Abdulaziz Üniversitesi, Matematik Bölümü, Suudi Arabistan
[3] Leeds Üniversitesi, Kanser ve Patoloji Enstitüsü, Leeds, İngiltere
[4] Leeds Üniversitesi, İstatistik Bölümü, Leeds Üniversitesi, Leeds, İngiltere

Amaç: Değişken sayısının gözlem sayısından fazla olduğu çok-boyutlu veriler (high-dimensional data) üzerinde modelleme yapmanın bazı zorlukları vardır. Ayrıca, bu verilerde değişkenler arasında yüksek korelasyon durumları gözlenir. Bu tür veriler için, "seyrek (sparse)" çözümler önerilir. Bu çözümler ile elde edilen modeller daha kolay yorumlanır. Bu çalışmanın amacı, rastgele etkilerin Gamma dağıldığı varsayımı altında, lojistik regresyon yöntemi kullanılarak seyrek çözüm elde etmektir.

Yöntem: Seyrek çözüm elde etmek için önerilen yöntemlerden biri Lasso cezasıdır. Ancak, bu yaklaşımın ilişkili değişkenler bağlamında bazı dezavantajları vardır. Bu çalışmada, Lasso cezasına alternatif olarak, rastgele etkilerin gamma dağıldığı varsayımı altında Lee ve Oh (2009) tarafından önerilen, Hiyerarşik olabilirlik (HL) yöntemi kullanıldı. Bu yöntem, Ridge ve Lasso cezalarının özel bir hali olarak kabul edilebilir. Ayrıca, HL cezasını Ridge cezası ile genişleterek, Elastik net cezasına benzeyen "HL net" adında bir ceza yöntemi önerildi.

Bulgular: Çalışmada Ridge, Lasso, HL, Elastik net ve HL net cezaları benzetim çalışmasında ve gerçek bir veri seti olan Akciğer hastalarının patolojik alt tiplerinin tahmininde kullanıldı ve sonuçları karşılaştırılarak tartışıldı. Gerçek veri seti, Leeds Eğitim Hastanesi'nde tedavi gören 76 akciğer kanseri hastasından oluşmaktadır. Benzetim çalışmasında ise, iki farklı senaryo oluşturuldu. Her iki senaryo için, ceza yöntemlerini içeren model performansları değerlendirildi. Gerçek veri setinde ise, faklı λ değerleri için çapraz doğrulama ile sınıflandırma hatası hesaplandı. Bulgulara göre, gerçek ve benzetim veri setinde HL cezası, Lasso cezasından daha seyrek tahminler verdi. HLnet ve Elastik net cezalarının ise gerçek veri setinde sınıflandırma hatalarının düşük çıktığı ve en iyi tahmin performansına sahip olduğu görüldü.

Sonuç: Ridge cezasına HL cezasının eklenmesi ile HL cezasının hassasiyeti benzetim çalışmasında ve gerçek verilerde artmıştır. Dolayısıyla (çok) seyrek çözüm istendiğinde HLnet cezası kullanılması önerilir.

Anahtar Sözcükler: Lojistik Regresyon, Hiyerarşik olabilirlik (HL) yöntemi, Seyrek Çözüm, Rastgele Etkiler.

Kaynaklar:

1. Y. Lee and J.A. Nelder, *Hierarchical generalized linear models*, J. R. Stat. Soc. Ser. B (Methodol.)58 (1996), pp. 619–678.
2. Y. Lee, J. Nelder, and Y. Pawitan, *Generalized Linear Models with Random Effects: Unified Analysis Via H-likelihood*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, 2006.
3. A. Gusnanto, H.M. Wood, Y. Pawitan, P. Rabbitts, and S. Berri, *Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data*, Bioinformatics 28 (2012), pp. 40–47. doi:10.1093/bioinformatics/btr593.
4. Y. Lee and H.S. Oh, *A new sparse variable selection via random-effect model*, J. Multivar. Anal.125 (2014), pp. 89–99. Available at http://www.sciencedirect.com/science/article/pii/S0047259X13002583.
5. Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, 2013. Available at https://books.google.co.uk/books?id=8T8fAQAAQBAJ.
6. Arief Gusnanto & Yudi Pawitan (2015) Sparse alternatives to ridge regression: a random effects approach, Journal of Applied Statistics, 42:1, 12-26, DOI:10.1080/02664763.2014.929640
7. Arief Gusnanto, Peter Tcherveniakov, Farag Shuweihdi, Manar Samman, Pamela Rabbitts and Henry M. Wood. *Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data.* Bioinformatics, 31(16), 2015, 2713–2720 doi: 10.1093/bioinformatics/btv191

S30

Mikrobiyom Verilerinde Kullanılan İstatistiksel Analizler: Obezite İlişkili Mikrobiyom Sekans Verisi Üzerinde bir Uygulama (Sözlü)

Selen Yılmaz Işıkhan[1], Ceren Özkul[2]

[1]*Hacettepe Üniversitesi Sosyal Bilimler Meslek Yüksekokulu, Ankara*
[2]*Hacettepe Üniversitesi Eczacılık Fakültesi, Farmasötik Mikrobiyoloji Anabilim Dalı, Ankara*
e-mail: seleny@hacettepe.edu.tr

Özet

Giriş: Kültürden bağımsız, yüksek çıktılı sekans yöntemlerinin yaygın kullanım alanı bulmasıyla bir ekosistemdeki mikroorganizmaları, gen ve gen ürünlerini, metabolik aktivitelerini kapsayan mikrobiyom çalışmaları her geçen yıl artmaktadır. Bu çalışmalardan elde edilen verilerin işlenmesi ve popülasyon analizleri ile ilgili çok çeşitli yaklaşımlar bulunmaktadır. Bu amaçla bu bildiride mikrobiyom çalışmalarında sık kullanılan bir yaklaşım olan 16S rRNA amplikon sekansı ile elde edilecek verilerin işlenmesi ve mikrobiyal popülasyonun istatistiksel analizleri ile ilgili R yazılımında bir akış şeması sunulmuştur.

Yöntem: Sunulan analiz şeması açık erişimli veri tabanlarından ham veri dosyaları indirilip işlenmiş "TwinsUK mikrobiyota" projesine dahil edilmiş olan obez, aşırı kilolu ve normal kilolu gönüllülerin rastgele seçilen verilerini içermektedir (Goodrich ve diğ. 2014). Çalışmanın ham sekans okumaları European Nucleotide Archive (ENA) ERP006339 ve ERP006342 numaraları ile kayıtlıdır. Alfa çeşitliliğin gruplar arasındaki değişimi observed-OTU, Shannon diversity, Simpson diversity ve Chao diversity metrikleri incelenerek değerlendirilmiştir. Beta çeşitliliği için OTU'lerin hem varlığı/ yokluğu hem de bolluğunu dikkate alan OTU'ye dayalı Bray-Curtis metriği kullanılarak nMDS (stres) değerleri hesaplanmıştır. PCoA kullanarak filogenetik uzaklığa dayalı ağırlıklı ve ağırlıksız UniFrac metrikleri elde edilerek saçılım grafikleri ile gösterilmiştir. Kilo grupları arasında taksa bolluk karşılaştırmaları Kruskal Wallis testi ile; ikili karşılaştırmalar ise (eşleştirilmemiş) Wilcoxon testi ile gerçekleştirilmiştir. Klasik testlere ek olarak, grupları tanımlayan en anlamlı taksa'ları seçmeyi sağlayan doğrusal diskriminant analizi etki büyüklüğü (LEfSe) analizi kullanılmıştır.

Bulgular: Alfa çeşitlilik metrikleri kilo grupları arasında benzer bulunmuştur ($p \geq 0.05$). Beta çeşitliliği Bray Curtis metriği de kilo grupları ile ilişkili bulunmamıştır ($F=1.01$, $p=0.343$). Permütasyon testi sonucunda da gruplar arası fark anlamlı değildir ($Nperm=999$, $F=0.664$, $p=0.534$). Normal, obez ve aşırı kilolu gruplar arasında Firmicutes ve Proteobacteria dağılışları açısından anlamlı fark bulunmamıştır (p değerleri: 0.47, 0.37). Ancak Euryarchaeota bakterisi üç grup arasında istatistiksel olarak anlamlı farklılık göstermiştir ($\chi^2=6.91$, $p=0.031$). Pairwise Wilcoxon testlerine (göre obez grubun ortanca bolluğu ($\tilde{x}=0.00$) aşırı kilolu gruba ($\tilde{x}=0.01$) kıyasla daha küçüktür ($p=0.029$). Lefse analizi ile $p<0.05$ ve LDA >2 etki büyüklüğü için, aşırı kilolularda en fazla farksal bolluğa sahip cins *Subdoligranulum* olarak bulunmuştur.

Sonuç: Son yıllarda mikrobiyom verilerinin kompozisyonel (birleşimsel-bağımlı) yapısını analiz etmenin her adımında Aitchison'un önerdiği Compositional Data Analysis (CoDA) metotları tercih edilmelidir. Grupların bakteriyel bolluk dağılışlarını toplu olarak görüntülemek amacıyla heatmap'ler ve boxplot'lardan yararlanılabilir. Klasik istatistiksel analizlere kıyasla tüm taksa topluluğunda grupları en fazla ayırt etmeyi sağlayan bakterileri seçebilen LEfSe analizi iyi bir alternatiftir. Alfa ve beta çeşitlilikleri açısından yapılan gruplar arası karşılaştırmalarda fark bulunmaması literatürdeki mikrobiyotayı obeziteye dayalı kilo gruplarında inceleyen bazı çalışmalar ile uyumludur (Kitten ve ark, Mammadova ve ark.).

Anahtar Sözcükler: Mikrobiyom sekans verisi, alfa çeşitlilik, beta çeşitlilik, taksonomik göreli bolluk karşılaştırması, LEfSe analizi.

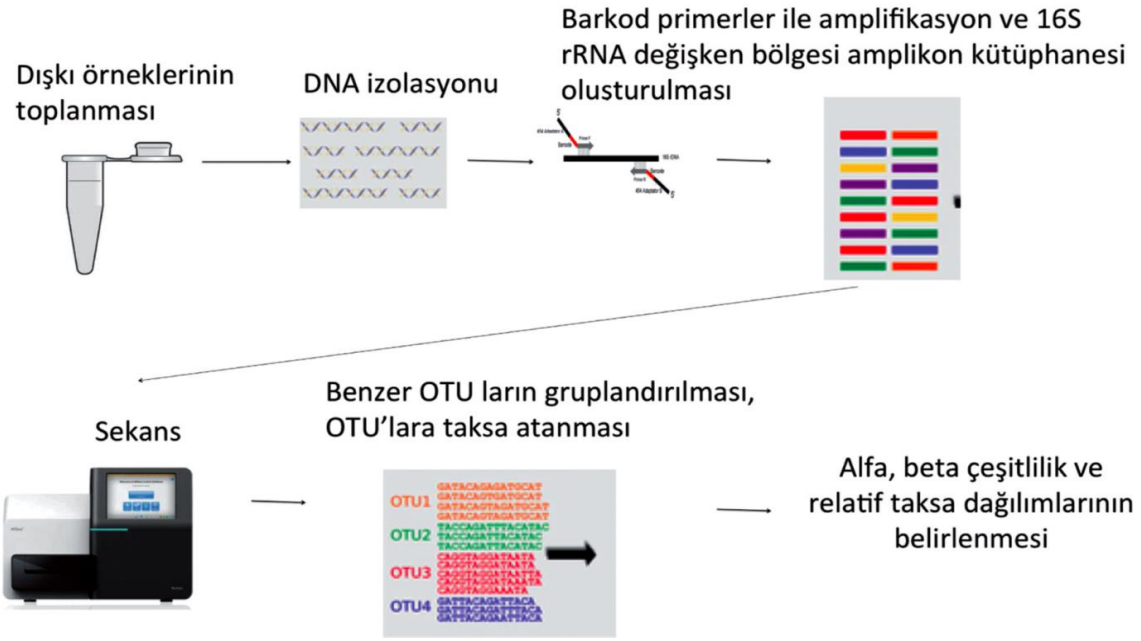
Giriş

Yüksek çıktılı yeni nesil dizilemenin ortaya çıkışı, genetik araştırmalarda devrim yaratmıştır; 2003 yılında insan genom projesinin tamamlanmasından bu yana, artık tüm insan genomu, belirli bir maliyetle ve belirli bir zaman içinde dizilenebilmektedir. Son 5 ila 10 yıl içinde, bu teknoloji, bir ekosistemde yaşayan tüm mikroorganizmaların kolektif genomlarını ve gen ürünlerini tanımlamak amacıyla 'mikrobiyom' çalışmalarında kullanılmaya başlanmıştır. İnsan mikrobiyomu, besinlerin sindirimi ve bağışıklık sisteminin modülasyonu gibi çok sayıda temel

işlevde yer alır ve mikrobiyom kompozisyonundaki değişikliklerin insan sağlığı üzerinde önemli etkileri olduğu gösterilmiştir [1]. Mikrobiyom verilerinin analizi, yüksek boyutlu yapılandırılmış çok değişkenli seyrek verileri içermesinden, operasyonel taksonomik ünitelerin (OTU) bileşimsel yapısından, sayım verisi analizinin olağan zorluklarından, yani çarpık dağılım, sıfır yığılmalı ve aşırı yaygın dağılımla karşı karşıya kalınmasından dolayı zorlayıcıdır [2-4]. Bu çalışmada mikrobiyom analizi için en yaygın kullanılan istatistiksel teknikleri, özellikle R paketlerinde uygulanabilen metot ve grafikler yardımıyla özetlemek amaçlanmıştır. Analizin her aşaması için yöntemlere çözüm fonksiyonlarının mikrobiyom çalışmasına nasıl uyarlanacağı açıklanmaktadır.

Mikrobiyom çalışmalarında temel akış şeması

Mikrobiyom çalışmalarında temel basamaklar örneğin eldesi, DNA izolasyonu, 16S rRNA'da seçilecek değişken bölgenin çoğaltılarak amplikon kütüphanelerinin oluşturulmasını, purifikasyon ve sekans basamaklarını içermektedir. Özet bir akış şeması Şekil 1'de verilmiştir.



Şekil 1. Mikrobiyota çalışmalarında özet akış şeması [5].

Örnekler uygun koşullar altında alındıktan sonra uygun izolasyon kitleri ile DNA izolasyonu gerçekleştirilmektedir. DNA izolasyonu sonrası amplikon sekansı için 16S rRNA'nın değişken bölgelerinden sıklıkla V3-V4 bölgeleri kullanılmaktadır. İlgili bölgelerin barkod primerler kullanılarak çoğaltılması, örneklerin havuzlanması sonrası Illumina Miseq gibi sekans platformlarında çift yönlü dizilemeleri yapılır. Kalite kontrol basamakları sonrası, sekansların ne kadar benzer olduğuna bağlı olarak QIIME gibi biyoinformatik platformlar kullanılarak operasyonel taksonomik birimlere (OTU'lar) kümelenir. Mikrobiyom dizilemede, bakteriler genetik benzerliklerine bağlı olarak kümelenir ve daha sonra bir cins, tür vb. olarak tanımlanır. %97 benzerlikte gruplama tür düzeyinde tanımlamaya büyük oranda izin verir. OTU'lar kümelendikten sonra her bir OTU ID'si için GreenGenes gibi veritabanları kullanılarak taksa atanır. Bu komut sonrası OTU tabloları sıklıkla 'biom' formatında elde edilir [1, 5].

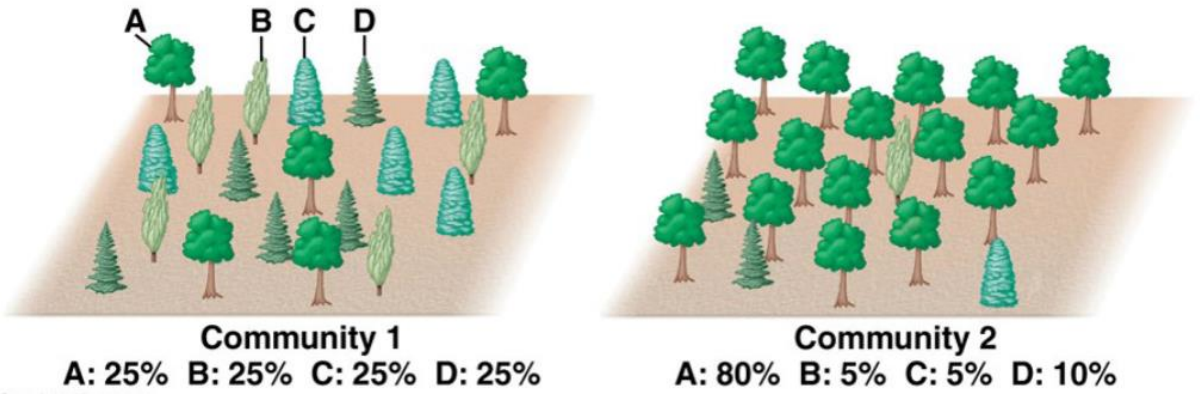
Sıralama verilerini işlemek için çeşitli yazılım paketleri mevcuttur. En sık kullanılanlar Qiime, ve Mothur'dur. QIIME, komut satırı girişlerini kullanır. Ön işlemler sonrası elde edilen özellik tabloları aşağıda detaylandırıldığı şekilde işlenerek mikrobiyal populasyon analizleri için temel çıktılar (alfa, beta çeşitlilik, taksa baskınlıkları) elde edilir.

Yöntem

Mikrobiyota çalışmalarında sıklıkla kullanılan alfa ve beta çeşitlilik analizleri ile iki veya daha fazla örneklem grubu arasındaki mikrobiyal bileşim farklılıklarının karşılaştırmasında kullanılabilecek istatistiksel metotlar sırayla açıklanmaktadır.

Alfa Çeşitlilik

Alfa çeşitliliği bir örnek içindeki çeşitliliği değerlendirir. Bu, her bir örnekte kaç farklı tür (OTU) olduğu (richness, zenginlik) ve bunların ne kadar eşit dağıldığı (evenness, düzgünlük), ki bunlar birlikte çeşitliliktir. Şekil 2 zenginliği ve çeşitliliği göstermektedir. Her iki orman da aynı zenginliğe (4 ağaç türü) sahiptir, ancak Topluluk 1, 4 türün çok daha eşit dağılımına sahipken Topluluk 2'de ağaç türleri A'nın hakimiyeti altındadır. Bu, Topluluk 1'i Topluluk 2'den daha çeşitli kılar.



Şekil 2. Alfa Çeşitlilik ve homojenlik değerlendirmesi

Alfa çeşitliliği hesaplamak için Qiime ve R'ın vegan paketi kullanılabilir. En sık kullanılan alfa çeşitlilik metrikleri Shannon's diversity, Simpson's diversity, Chao's richness ve ACE's richness'tır. Her gruptaki örneklerin alfa çeşitlilik değerleri, kutu grafikleri kullanılarak görsel olarak karşılaştırılabilir. Gruplar arasındaki alfa çeşitlilik farklılıkları, veriler kabaca normal dağıldığında Varyans Analizi (ANOVA), t-testi ya da genelleştirilmiş lineer modeller ile; veri normal dağılıma uymadığında ise Kruskal Wallis, Mann-Whitney U testi, Wilcoxon sıra toplamları testi ya da farklı dağılımlı genelleştirilmiş lineer modeller kullanılarak istatistiksel olarak değerlendirilebilir [6, 7]. Alfa çeşitlilik metriklerinin normal dağılım gösterip göstermediği hist(meta$shannon) ve shapiro.test(meta$shannon) gibi fonksiyonlarla görsel ve teste dayalı olarak değerlendirilebilir.

Beta Çeşitlilik

Beta çeşitlilik, örnekler arasında mikrobiyomdaki farklılıkları değerlendirir. Dolayısıyla, her numunenin birden fazla değeri vardır. Bazı metrikler bolluğu hesaba katar (yani çeşitlilik: Bray-Curtis, ağırlıklı UniFrac) ve bazıları yalnızca varlık-yokluk temelinde hesaplar (yani zenginlik: Jaccard, ağırlıksız UniFrac).

Beta çeşitliliği görsel olarak temsil etmek için temel koordinatlar analizi (PCoA), metrik olmayan çok boyutlu ölçekleme (NMDS) veya kısıtlı ana koordinatlar analizi (CPCoA) gibi boyut indirgeme yöntemleriyle birleştirilir [6, 7]. Bu analizler R Vegan paketinde yapılabilir ve saçılım grafiği şeklinde görselleştirilebilirler. Bir nMDS grafiğindeki her sembol, o örneğin toplam mikrobiyal topluluğunu temsil eder. Birbirine daha yakın olan semboller daha fazla benzer mikrobiyotaya sahipken, daha uzak olan semboller daha az benzerliğe sahiptir.

p1= (p11,…,p1k) ve p2= (p21,…,p2k) iki farklı örneğin mikrobiyom nispi bolluğunu göstersin. Bray-Curtis şu şekilde tanımlanır:

$$d_{BC}(p_1, p_2) = \frac{\sum_{i=1}^{k} |p_{1i} - p_{2i}|}{\sum_{i=1}^{k} (p_{1i} + p_{2i})}.$$

Beta çeşitlilik endeksleri arasındaki istatistiksel farklar, veganda adonis() işleviyle permütasyonlu çok değişkenli varyans analizi (PERMANOVA) kullanılarak hesaplanabilir.

Filogenetiğe Dayalı Metrikler

Beta çeşitlilik metriklerinin en yaygını UniFrac'tır. UniFrac'ın Bray-Curtis veya Jaccard'a göre gücü, mikrobiyotada bulunan türlerin filogenetik ilişkilerini hesaba katmasıdır. Tüm UniFrac varyantları, iki örnek

arasındaki mesafeyi hesaplamak için merkezi bilgi olarak taksonlar arasındaki ilişkinin filogenetik ağacını kullanır.

Taksonomik kompozisyon, bir mikrobiyal toplulukta bulunan mikrobiyota elemanlarını tanımlar. Veri görselleştirilirken genellikle phylum veya genus düzeyinde gösterilir. Taksa ve örneklem meta verisi arasındaki ilişkileri ortaya çıkarmak için korelasyon analizi kullanılır. Ağ analizi, özelliklerin birlikte oluşumunu bütünsel bir bakış açısıyla araştırır. Korelasyon katsayıları ve önemli P değerleri, R'deki cor.test() işlevi veya SparCC (kompozisyonel veriler için seyrek korelasyonlar) paketi gibi bileşimsel veriler için uygun olan daha sağlam araçlar kullanılarak hesaplanabilir [8]. Ağlar ayrıca R kütüphanesi igraph kullanılarak görselleştirilebilir ve analiz edilebilir [9].

Makine Öğrenmesi ve Veri Madenciliği Yöntemleri

Makine öğrenimi, verilerden öğrenen, kalıpları tanımlayan ve kararlar veren bir yapay zeka dalıdır. Mikrobiyom araştırmalarında, taksonomik sınıflandırma, beta-çeşitlilik analizi, gruplama ve belirli özelliklerin kompozisyon analizi için makine öğrenimi kullanılır. Yaygın olarak kullanılan makine öğrenimi yöntemleri biyolojik belirteç bolluğunda deneysel duruma bağlı değişiklikleri göstermek için regresyon analizini, biyomarker'ı seçerek grupları sınıflandırmak için random forest (RF), Destek vektör makinesi (DVR), Adaboost ve deep learning'i içerir. RF ve DVR'yi hesaplamak için R'ın caret paketi ve randomForest paketi de kullanılabilir.

Modeli eğitmek ve oluşturulan modelin performansını değerlendirmek için OTU bolluk tablosunun %70'ini eğitim seti olarak ve kalan %30'unu ise test seti olarak kullanılır. Makine öğrenimi modelinin performansını değerlendirmek için, Alıcı işletim karakteristiği (ROC) eğrisi modelimizin test setimizle ne kadar iyi performans gösterdiğini gösterir.

Farksal Bolluk Analizi

R'ın BiomMiner hesaplama aracı 16S rRNA veya tüm genom verisetinde farklı bolluktaki özellikleri tanımlamak için metodoloji gerçekleştirir. Bunun için Metastats, LEfSe ve mothur v.1.34'te bulunan Kruskal–Wallis testi gibi köklü yöntemler kullanılır. Metastats, Fisher'in kesin testini gerçekleştirir ve iki grup arasında farklı olan özelliklerin bir listesini sağlamak için p değerini hesaplar.

Çok Değişkenli Farksal Bolluk Testi

Çok değişkenli diferansiyel bolluk testi, iki veya daha fazla numune grubu arasındaki mikrobiyal bileşimdeki farklılıkların global bir testini ifade eder. Kruskal-Wallis (sıra sayılarında tek yönlü ANOVA), ikiden fazla topluluk arasındaki özelliklerin aynı dağılımdan kaynaklanıp kaynaklanmadığını test eden parametrik olmayan bir yöntemdir. Uzaklık matrisini kullanan Permütasyonlu Çok Değişkenli Varyans Analizi (PERMANOVA) en yaygın kullanılan türüdür. Gruplar arasında bileşimde fark olmadığına dair yokluk hipotezi farklı örnek gruplarının aynı kütle merkezine sahip olması koşuluyla formüle edilmiştir. Farklılıklara dayalı Vegan R paketinin "adonis" işleviyle uygulanan ANOVA ile, benzerliklerin analizi ise Vegan R paketinin "anosim" işlevi ile gerçekleştirilir [10].

Tek Değişkenli Farksal Bolluk Testi

Mikrobiyom bileşiminde önemli küresel farklılıklar olduğunda bu farklılığa hangi belirli taksaların yol açtığı sorgulanır. Wilcoxon sıra toplamı testi veya Kruskal-Wallis testi gibi parametrik olmayan testler uygulanabilir. Ancak, klasik tek değişkenli yaklaşımların özellikle mikrobiyom verilerinin kompozisyon yapısından etkilendiğini ve sonuçlarının büyük yanlış pozitif oranlara sebep olduğu bildirilmiştir.

Mikrobiyom verilerinin bileşimsel yapısını açıklayabilen iki CoDA yöntemi ANCOM ve ALDEx2'dir.

- ANCOM tüm değişken çiftlerinin log oranı, ortalamalardaki farklılıklar için test edilir.
- ALDEx2 algoritması sayımlardan çok değişkenli bolluk dağılımını çıkarmak için bir Dirichlet-çok terimli model kullanır [11].

LEfSe Analizi

LEfSe (Doğrusal diskriminant analizi Etki Boyutu), etki düzeyini tanımlamak için bir Kruskal-Wallis ve bir Wilcoxon sıra toplamı testini, istatistiksel anlamlılık için ise bir Doğrusal Ayırma Analizi (LDA) uygulayarak

topluluklar arasındaki farklılıkları açıklama olasılığı en yüksek olan özellikleri (OTU) seçer. Mikrobiyom çalışmalarında biyobelirteç saptamak için kullanılan bir yaklaşımdır [12]. LDA modeli grup etiketini bağımlı değişken olarak, önceki adımda seçilen taksonların gözlenen bolluğu ve demografik değişkenleri ise bağımsız değişkenler olarak dikkate alarak oluşturulur. Spesifik olarak, ilgilenilen sınıfa göre anlamlı diferansiyel bolluğu olan özellikleri tespit etmek için ilk önce parametrik olmayan faktöriyel Kruskal-Wallis testi kullanılır; biyolojik tutarlılık daha sonra (eşleştirilmemiş) Wilcoxon sıra toplamları testi kullanılarak alt sınıflar arasında bir dizi ikili test kullanılarak araştırılır. Son bir adım olarak, LEfSe, diferansiyel bolluğa sahip her özelliğin etki büyüklüğünü tahmin etmek ve araştırmacı tarafından istenirse, boyut küçültme gerçekleştirmek için LDA'yı kullanır.

Heatmap

Bir veri kümesindeki takson sayısı çok arttıkça, verilerin tüm öğelerini etkin bir şekilde görüntüleme yeteneği zorlaşır ve Heatmap bu anlamda yardımcı bir araçtır. Heatmap'i daha okunabilir hale getirmek için en bol bulunan ilk 20 OTU'ye taksonlar atanarak R'ın plot_heatmap fonksiyonundan yararlanılabilir.

Veriseti

Sunulan analiz şeması açık erişimli veri tabanlarından ham veri dosyaları indirilip işlenmiş "TwinsUK mikrobiyota" projesine dahil edilmiş olan obez, aşırı kilolu kilolu ve zayıf gönüllülerin verilerini içermektedir (Goodrich et al. 2014 - doi: 10.1016/j.cell.2014.09.053). Çalışmanın ham sekans okumaları European Nucleotide Archive (ENA) ERP006339 ve ERP006342 numaraları ile kayıtlıdır [13].

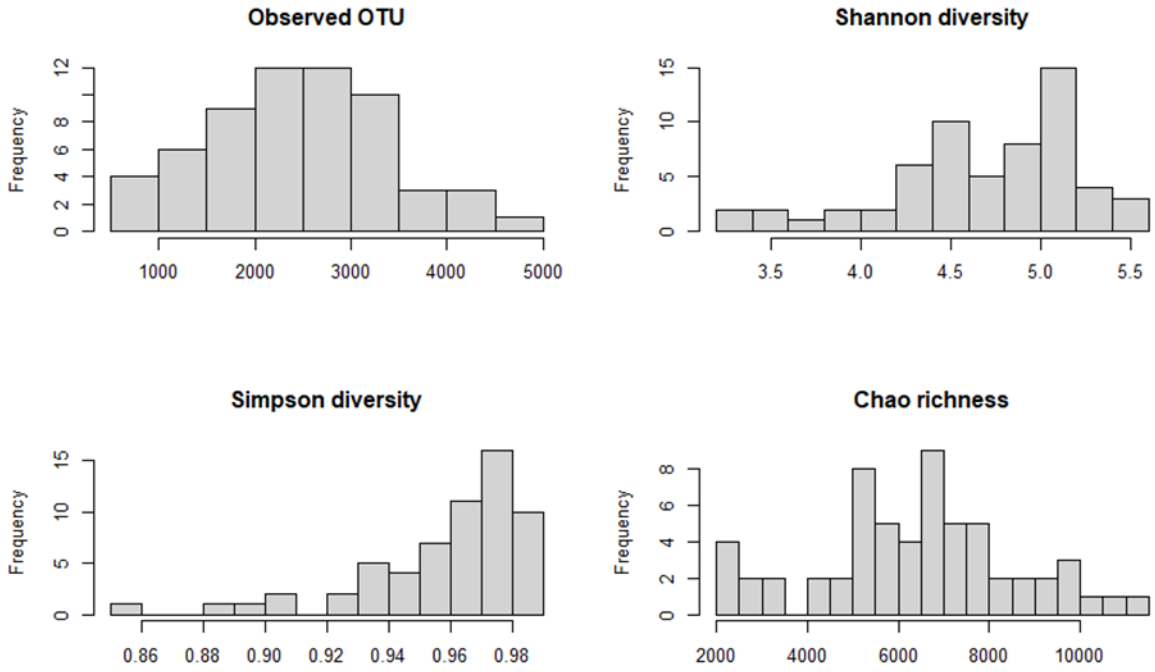
Verisetinde 416 ikiz çift dahil olmak üzere TwinsUK popülasyonundan elde edilen> 1.000 dışkı örneği kullanılmıştır. 977 kişiden 1.081 dışkı örneği alınmış: 171 MZ ve 245 DZ ikiz çifti, 2'si bilinmeyen zigositeye sahip ikiz çiftlerinden ve 143 örnek ise bir ikizlik içindeki sadece tek ikizden alınmıştır. Deneklerin çoğu, 23 ila 86 yaşları arasında değişen kadındı (ortalama yaş: 60.6 ± 0.3 yıl). Deneklerin ortalama BKİ (Beden kitle indeksi)'si 26.25 (± 0.16) idi: dağılışları ise 433 denek düşük ila normal BKİ'ye sahipti (<25), 322 kişi aşırı kilolu BKİ'ye sahipti (25-30), 183 kişi obezdi (>30) ve 39 kişi mevcut BKİ durumunun bilinmediği kişilerdi. Çalışmamızda mikrobiyom analizlerini göstermek amacıyla her bir gruptan rastgele 20'şer örnek çekilmiştir.

Bulgular

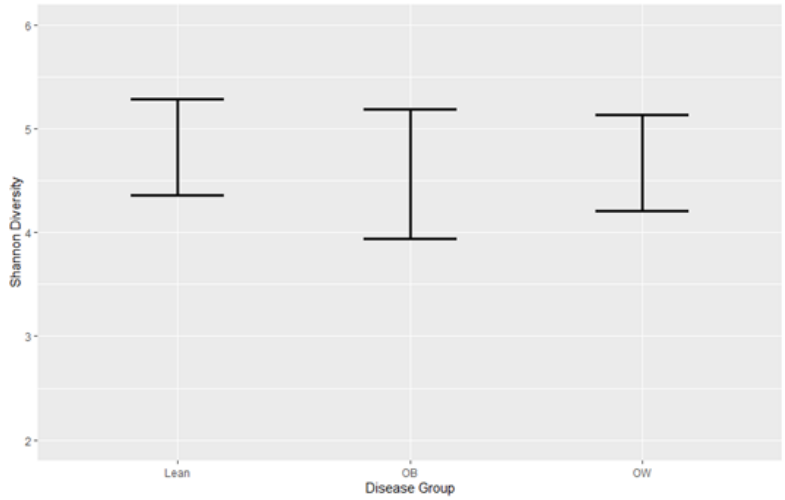
Alfa Çeşitlilik değerlendirmesi

Alfa çeşitliliği ve zenginliğin genel değerlendirmesi için ilk olarak 4 farklı metriği eşzamanlı gözlemlemek için histogram grafikleri oluşturulmuştur (Şekil 3). Grafiklere göre gözlenen OTU, Shannon ve chao zenginlik metriklerinin kabaca normal dağıldığı görülmektedir. Kolmogorov-Smirnov testi de sadece 1/simpson ölçüsünün normal dağılmadığını doğrulamıştır (p=0.013).

Üç kilo grubu arasında yapılan karşılaştırmada; Shannon çeşitliliği zayıf grupta diğer kilo gruplarına göre daha yüksek gibi görünmesine rağmen grup dağılışları arasında anlamlı fark bulunmamıştır (F=1.215, p=0.304). Diğer ölçüler açısından da gruplar arasında fark yoktur (p≥0.05, Şekil 4).



Şekil 3. Alfa çeşitlilik metriklerinin dağılışları



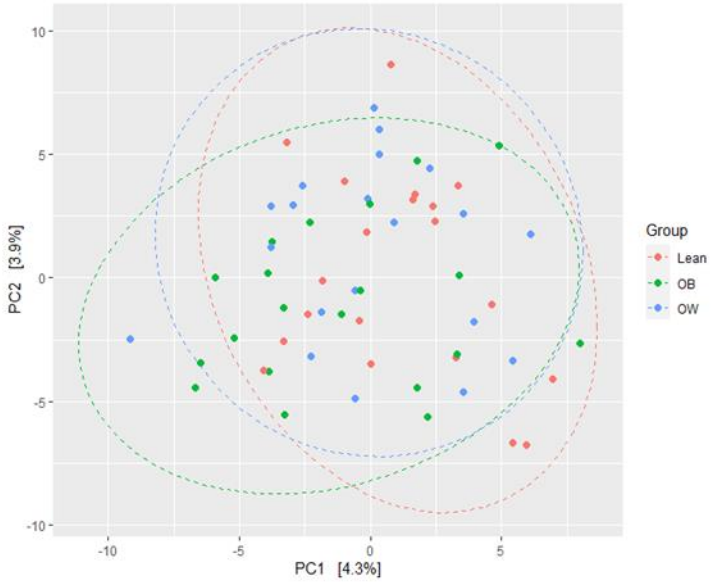
Şekil 4. Kilo gruplarının Shannon çeşitlilik metriği dağılışı (Lean: Zayıf, OB: Obez, OW: Aşırı kilolu)

Beta Çeşitlilik Sonuçları

Beta çeşitlilik için ilk olarak; OTU'lerin hem varlığı/ yokluğu hem de bolluğunu dikkate alan OTU'ye dayalı Bray-Curtis metriği kullanılarak nMDS değerleri hesaplanmıştır. K=2 ve k=3 boyutlu uzayda BC uzaklığına dayalı çözümler için yakınsama elde edilememiştir. Buna rağmen stress değerleri k=2 için 0.24 ve k=3 boyutta ise 0.19 bulunmuştur. Aitchison uzaklığını kullanarak beta çeşitlilik ordinasyonu elde edilmiştir. Özdeğerler ve açıkladığı varyans oranları hesaplanmıştır. İlk 5 bileşenin açıklanan varyansları aşağıdaki gibidir:

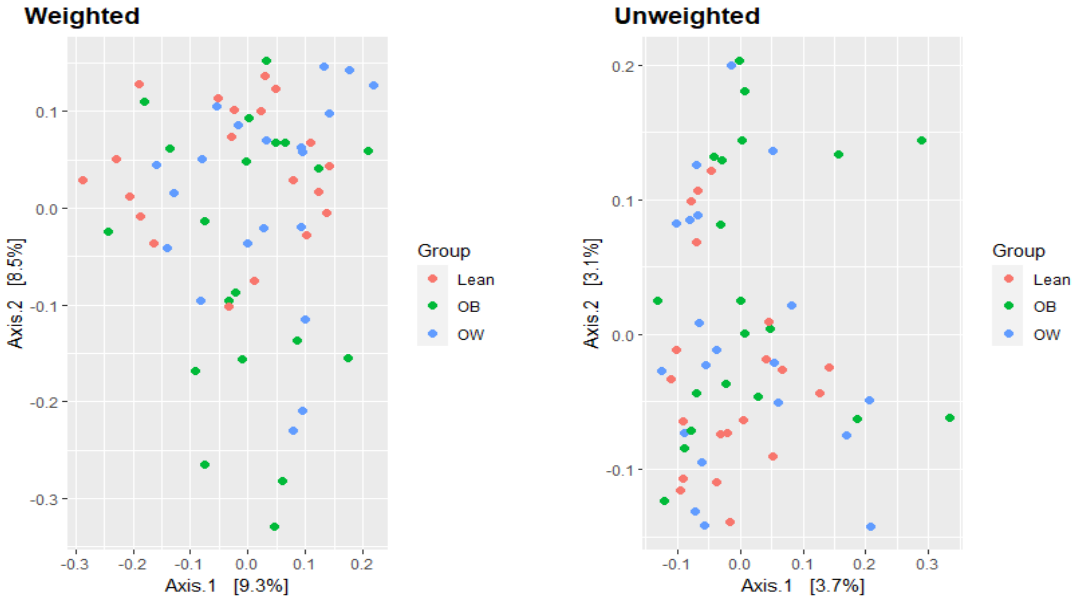
PC1	PC2	PC3	PC4	PC5
0.04294397	0.03930212	0.03558859	0.02990568	0.02631134

Kilo gruplarına göre merkez konumunda fark olmadığına dair yokluk hipotezi kabul edilmiştir (F=1.01, p=0.343, Şekil 5). Gruplar arası açıklanan varyans oranı da oldukça küçüktür (R^2=0.034). Yine permütasyon testi sonucunda da (Nperm=999, F=0.664, p=0.534) gruplar arası fark anlamlı değildir.



Şekil 5. Bray-Curtis'e dayalı PcoA

Son olarak PCoA kullanarak filogenetiğe dayalı ağırlıklı ve ağırlıksız UniFrac metrikleri hesaplanmıştır (Şekil 6). PCoA, Öklid dışı metrikler için temel bileşenler analizi olarak düşünülebilir.



Şekil 6. Ağırlıklı ve ağırlıksız UniFrac metrikleri için PCoA

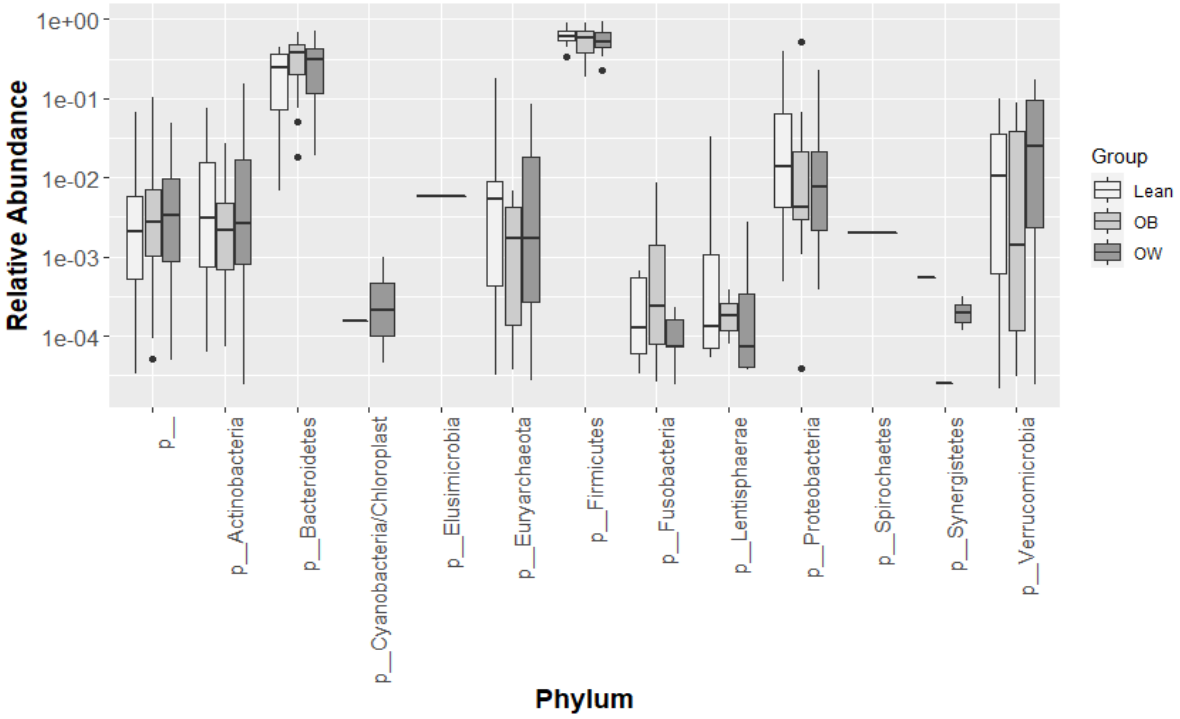
Relatif Taksa Baskınlıkları

Phylum düzeyinde Firmicutes, Proteobacteria ve Euryarchaeota'nın kilo grupları arasındaki dağılışları Kruskal Wallis ile test edilmiştir. Zayıf, obez ve aşırı kilolu gruplar arasında Firmicutes ve Proteobacteria dağılışları açısından anlamlı fark bulunmamıştır (p değeri sırasıyla 0.47, 0.37). Ancak Euryarchaeota bakterisi üç hastalık grubu arasında istatistiksel olarak anlamlı farklılık göstermiştir (χ^2=6.91, p=0.031). Pairwise Wilcoxon (ikili karşılaştırma) testlerine göre bu farklılık obez ve aşırı kilolu gruplar arasındaki farktan kaynaklanmıştır (p=0.029). Obez grubun ortanca bolluğu aşırı kilolu gruba kıyasla daha düşüktür.

Euryarchaeota bakterisi için grupların tanımlayıcı istatistikleri

```
Zayıf
  vars  n  mean   sd median trimmed mad min  max range skew kurtosis   se
X1    1 20 0.01 0.04      0       0   0   0 0.18  0.18 3.73    12.79 0.01
------------------------------------------------------------------------------------------
Obez
  vars  n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 20    0  0      0       0   0   0 0.01  0.01 2.02     2.63  0
------------------------------------------------------------------------------------------
Aşırı kilolu
  vars  n  mean   sd median trimmed mad min  max range skew kurtosis se
X1    1 20 0.01 0.02      0    0.01   0   0 0.09  0.09 2.39     5.47  0
```

Gruplara göre taksa bolluklarının şekilsel sunumu ise ggplot2 paketi yardımıyla elde edilmiştir (Şekil 7).



Şekil 7. Phylum düzeyinde gruplar arası taksa bolluk karşılaştırması

Tablo 1. Grupların taksa bolluğunun Aldex2 Testi ile karşılaştırması

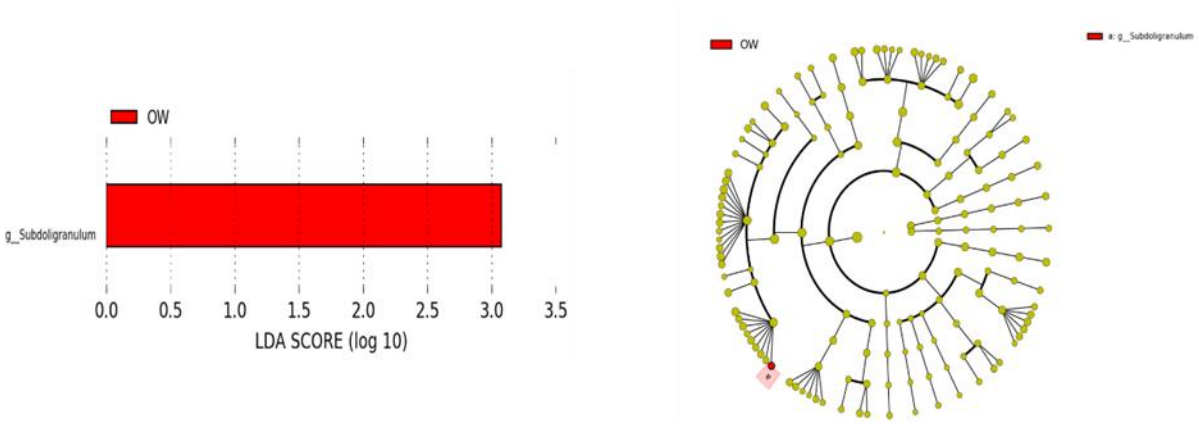
Genus Düzeyinde göreli bolluk karşılaştırma Bulguları

Genus Türü	Aldex2 Düzeltilmemiş KW P Değeri	Düzeltilmiş KW P Değeri
Lachnospiracea_incertae_sedis	0.032	0.723
Dorea	0.011	0.347

Bacteroides	0.013	0.501
Ruminococcus	0.037	0.773
Blautia	0.041	0.739
Faecalibacterium	0.036	0.744

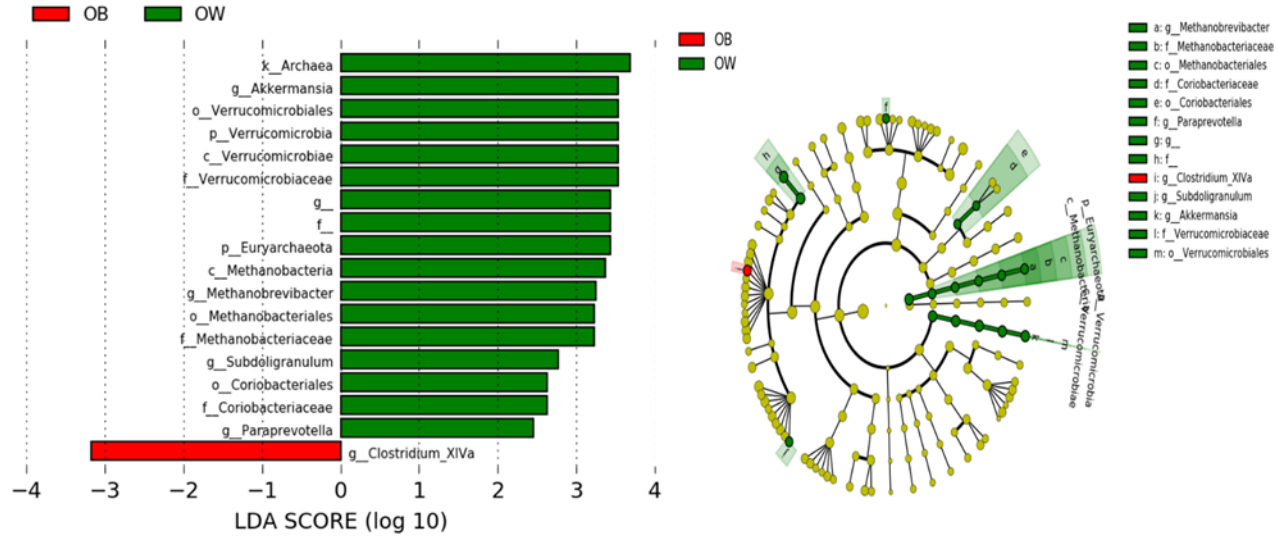
LEfSe Analizi Bulguları

%0.01 den düşük OTU ların filtrelenerek 3 Grup arasında yapılan Lefse analizi sonucunda; $p<0.05$ ve LDA >2 büyüklüğü için, aşırı kilolularda en fazla farksal bolluğa sahip taksa *Subdoligranulum* cinsidir.



Şekil 8. Zayıf, aşırı kilolu ve obez gruplar arası LEfSe analizleri

Grupların ikişerli karşılaştırmalarının alındığı LEfSe analizi sonuçlarına göre; yeşil olarak görülenler obeziteye anlamlı etkisi olan bakteriler, kırmızılar ise zayıf grup üzerinde anlamlı farksal bolluğa sahip bakteriler olmak üzere kladograma göre 11 adet tespit edilmiştir. Zayıf ve aşırı kilolular arası karşılaştırmada aşırı kilolu grup üzerinde anlamlı etkiye sahip bakteri Subdoligranulum iken zayıf grup üzerinde anlamlı bolluğa sahip taksalar ise 6 adet belirlenmiştir. Obez ve aşırı kilolu grup karşılaştırması sonucunda ise Clostridium_XIVa bakterisi obezite üzerinde anlamlı farksal bolluğa sahip iken aşırı kilolu grupta etkili taksalar ise 12 adet bulunmuştur (Şekil 9).



Şekil 9. Obez ve aşırı kilolu gruplar arası Lefse karşılaştırması

Tartışma ve Sonuç

Mikrobiyom sayım verileri genel olarak aşırı yaygın ve çok sayıda sıfıra sahip yapıdadır. Eğer sıfırlı hücreler için filtreleme, süzme gibi özel bir işlem uygulamıyor isek fazla sıfırların özelliklerini yakalamak ve sağa çarpık mikrobiyom verilerini modellemek için Zero-inflated Poisson (ZIP), Zero-inflated Negative Binomial (ZINB) veya hurdle modeli gibi bir sıfır yığılmalı modeller seçilmelidir [1, 14]. Mikrobiyom verilerinin kompozisyonel (bağımlı) yapısının göz ardı edilmesi de yapay korelasyonlar ile sonuçlanabilir. Son yıllarda mikrobiyom verilerinin kompozisyonel (birleşimsel-bağımlı) yapısını analiz etmenin her adımında (α ve β çeşitlilik analizleri, ordinasyon ve tek değişkenli ve çok değişkenli farksal bolluk testlerinde Aitchison'un önerdiği Compositional Data Analysis (CoDA) metotları tercih edilmelidir [1, 7]. Grupların bakteriyel bolluk dağılışlarını toplu olarak görüntüleme amacıyla heatmap'ler ve boxplot'lardan yararlanılabilir. Ancak klasik istatistiksel analizler için tüm taksa topluluğunda grupları en fazla ayırt etmeyi sağlayan bakterileri seçebilen Lefse testi iyi bir alternatiftir. Klasik istatistiksel metotlar ile Lefse testi bulgularındaki farklılıklar değişen parametre değerlerine göre benzetim deneyleri ile irdelenebilir. Alfa ve beta çeşitlilikleri açısından yapılan gruplar arası karşılaştırmalarda fark bulunamamış olması literatürdeki mikrobiotayı kilo gruplarında inceleyen bazı çalışmalar ile uyumludur [15, 16]. Kitten ve arkadaşları 14 diyabetli ve 23 kontrol hastası ile yürüttükleri çalışmada Shannon çeşitliliği ($p = 0.341$) ve beta çeşitliliğini Tip 2 diyabet ile ilişkili bulamamışlardır ($p = 0.201$).

Kaynaklar

1. Calle, M.L., *Statistical analysis of metagenomics data.* Genomics & informatics, 2019. **17**(1).
2. Fettweis, J.M., et al., *The vaginal microbiome and preterm birth.* Nature medicine, 2019. **25**(6): p. 1012-1021.
3. Zhou, W., et al., *Longitudinal multi-omics of host–microbe dynamics in prediabetes.* Nature, 2019. **569**(7758): p. 663-671.
4. Serrano, M.G., et al., *Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy.* Nature medicine, 2019. **25**(6): p. 1001-1011.
5. Özkul Koçak, C. and A. Yalınay, *Mikrobiyota Tanımlama Yöntemleri.* 2018.
6. Dill-McFarland, K.A., J.D. Breaker, and G. Suen, *Microbial succession in the gastrointestinal tract of dairy cows from 2 weeks to first lactation.* Scientific reports, 2017. **7**(1): p. 1-12.
7. Liu, Y.-X., et al., *A practical guide to amplicon and metagenomic analysis of microbiome data.* Protein & cell, 2021. **12**(5): p. 315.
8. Kurtz, Z.D., et al., *Sparse and compositionally robust inference of microbial ecological networks.* PLoS computational biology, 2015. **11**(5): p. e1004226.
9. Saito, R., et al., *A travel guide to Cytoscape plugins.* Nature methods, 2012. **9**(11): p. 1069-1076.
10. Plummer, E., et al., *A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data.* Journal of Proteomics & Bioinformatics, 2015. **8**(12): p. 283-291.
11. Gloor, G.B. and G. Reid, *Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data.* Canadian journal of microbiology, 2016. **62**(8): p. 692-703.
12. Segata, N., et al., *Metagenomic biomarker discovery and explanation.* Genome biology, 2011. **12**(6): p. 1-18.
13. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-799.
14. Xia, Y. and J. Sun, *Hypothesis testing and statistical analysis of microbiome.* Genes & diseases, 2017. **4**(3): p. 138-148.
15. Mammadova, G., et al., *Characterization of gut microbiota in polycystic ovary syndrome: findings from a lean population.* European Journal of Clinical Investigation, 2021. **51**(4): p. e13417.
16. Kitten, A.K., et al., *Gut microbiome differences among Mexican Americans with and without type 2 diabetes mellitus.* PloS one, 2021. **16**(5): p. e0251245.

S31

Mamografi Görüntülerinde Ön İşlemede Kullanılan Filtreleme Yöntemleri için Belirlenen Farklı Parametre Değerlerinin Sınıflama Başarısına Etkisi

Hanife Avcı[1], Jale Karakaya[1]

[1]*Hacettepe Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara*

Giriş: Meme kanseri kadınlarda en sık görülen kanser türlerinden biridir. Ancak görüntüleme yöntemleri ile düzenli tarama yapılarak erken teşhis edildiğinde tedavi edilebilir bir hastalıktır. Dijital mamografi meme kanseri tanısında kullanılan tarama yöntemidir. Hızlı ve doğru tanı koyabilmek için mamografi görüntülerinin yüksek kontrast ve çözünürlüğe sahip olması gerekir. Bilgisayar destekli tanı (CAD) modelleri, tıbbi görüntülerdeki lezyonlu alanları yorumlamaya ve teşhis etmeye yardımcı olan bilgisayar sistemleridir [1]. Görüntü işleme algoritmaları, CAD sistemleri ile mamografi görüntülerini iyileştirmek ve normal ve anormal dokuyu ayırt etmek için kullanılır. Çalışmanın amacı, mamografi görüntülerinin görünürlüğünü artırmak ve görüntülerdeki gürültüyü azaltmak için kullanılan çeşitli ön işleme yöntemlerinin parametre değerlerindeki değişikliklerin sınıflandırma performansına katkısını incelemektir.

Yöntem: Görüntüleme yöntemleri ile düzenli tarama yapılarak hastalıklara erken tanı konulması büyük önem taşımaktadır. Hızlı ve doğru tanı koyabilmek için görüntüleme yöntemleriyle elde edilen görüntülerin yüksek kontrast ve çözünürlüğe sahip olması ve etiket bilgisi içermemesi gerekmektedir. Bu nedenle bilgisayar destekli tanı modelleri, tıbbi görüntülerde lezyonlu bölgelerin yorumlanmasına ve tanı koyulmasına yardımcı olan bilgisayar sistemleridir. Bu nedenle, bu çalışmada, öncelikle görüntü ön işleme yöntemleriyle mamografi görüntüleri iyileştirilmiş, ön işleme yöntemlerinin farklı parametre değerlerine göre normal-benign-malign ve normal-anormal (benign ve malign) doku sınıflandırma başarıları incelenmiştir. Bunun için mini-MIAS [2] veri tabanından rastgele olarak seçilen 120 mamografi görüntüsü kullanılmıştır. Bu çalışmada ön işleme yöntemi olarak Gauss filtre, CLAHE ve Fast local Laplacian filtreleme yöntemleri uygulanmıştır. Gauss filtreleme yönteminin σ (sigma) parametresi -3<σ<3 aralığında değişim göstermektedir [3]. Laplacian filtreleme yöntemi için σ (sigma) ve α (alpha) parametre değerleri belirlenmektedir. Binary görüntüler için σ değeri [0, 1], farklı aralıkta görüntüler için σ değeri [a,b] aralığında olmalıdır. α değeri ise [0.01, 10] aralığındadır [4]. 1.Parametre değerleri Gauss filter için $\sigma = 3$, Laplacian filter σ=0.6 and α=0.6 ve 2. Parametre değerleri Gauss filter için $\sigma = 1$, Laplacian filter σ=2 ve α=2 olarak belirlenmiştir. Ön işleme adımından sonra k-ortalamalar kümeleme bölütleme yöntemi ile ilgili bölgeler (region of interest, ROI) elde edilmiştir. Gri seviye eş oluşum matrisi (Gray level co-occurence matrix, GLCM) ve Gri seviye çalışma uzunluğu matrisi (Gray level run length matrix, GLRLM) yöntemleri ile özellikler çıkarılmıştır. Korelasyon matrisi ile özellik seçimi yapıldıktan sonra Destek Vektör Makineleri (DVM), Rastgele Orman (RO), Yapay Sinir Ağları (YSA), Naive Bayes (NB), k-En Yakın Komşuluk (k-NN) ve Karar Ağaçları (KA) yöntemleri ile elde edilen özelliklerin performansları incelenmiştir.

Bulgular: 1. parametre değerlerine göre 2. parametre değerleri ile elde edilen özelliklerinden çoğunun tek başına sınıflama performanslarının daha yüksek olduğu görülmüştür. Seçilen 10 özelliğin birlikte değerlendirildiği MIAS veri tabanındaki mamografi görüntülerinin farklı parametre değerlerine göre, üçlü sınıf olarak (normal, benign and malign doku) sınıflandırma yöntemlerinin doğruluk oranları incelenmiştir. Bu sonuçlara göre Gauss bulanıklaştırma filtresinin parametre değeri küçüldükçe, Fast local Laplacian filtering parametre değerleri büyüdükçe daha yüksek doğruluk değerleri elde edilmiştir. Normal-anormal (benign+malign) sınıflamasında ise 2.parametre değerlerinin 1.parametre değerlerine göre hem doğruluk hem de eğri altında kalan (AUC) bakımından daha yüksek performans göstermiştir. Doğruluk ve AUC gibi ölçülere baktığımızda, 2.parametre değerlerine göre, DVM, RF, ANN ve NB sınıflandırma yöntemlerinin performansları k-NN ve DT yöntemine göre daha başarılıdır.

Sonuç: Özetle, mamografi görüntülerini iyileştirmek için kullanılan ön işleme yöntemlerinin farklı parametre değerlerinin sınıflandırma yöntemlerinin başarısını değiştirebileceği elde edilmiştir.

Başka bir deyişle, Gauss bulanıklık filtresinin parametre değeri azaldıkça ve Hızlı yerel Laplacian filtreleme parametre değerleri arttıkça sınıflandırma yöntemlerinin performansı arttı.

Anahtar Sözcükler: Bilgisayar destekli tanı, filtreleme yöntemleri, farklı parametre değerleri, veri madenciliği.

Kaynaklar

1. Scholl, I., Aach, T., Deserno, T. M., Torsten, K. 2010. Challenges of medical image processing. Comput. Sci. Res. Dev. 26, 5-13.
2. The Mini-MIAS Database of Mammograms (Internet). [Erişim tarihi 27 Eylül 2019]. Erişim adresi: http://peipa.essex.ac.uk/info/mias.html.
3. Besl, P. J., Jain, R. C. 1988. Segmentation through variable-order surface fitting. IEEE; 10(2), 167-192.
4. Bandyopadhyay, S. 2010. Pre-processing of Mammograms Images. International Journal of Engineering Science and Technology. 2(11), 6753-6758.

S32

Çevre Epidemiyolojisi Alanında Kullanılan Zaman Serisi Modellerinin Bibliyometrik ve Altmetrik Skorlarının Karşılaştırılması

Mehmet KARADAĞ

Hatay Mustafa Kemal Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Hatay, mkarad@gmail.com

Amaç: Hava kirliliği ve çöl fırtınası gibi olaylarla hastalık kaynaklı ölümlerin, zaman serisi regresyon modelleri kullanılarak ilişkilendirilmesi son yıllarda popüler olmuş bir alandır. Bu çalışmada bahsi geçen modelleri kullanan makalelere akademi (atıf) ile sosyal medyanın ilgisinin (Altmetrik İlgi Skorları (AİS)) karşılaştırılması ve modellerden hangisinin daha çok öne çıktığının incelenmesi amaçlandı.

Yöntem: Bu çalışmaya dâhil edilecek makalelerin incelenmesi, Thomson Reuters (Philadelphia, PA, ABD) tarafından sağlanan Web of Science Core Collection (WoSCC) aracılığıyla Science Citation Index Expanded (SCI-EXPANDED) veritabanı kullanılarak 29 Eylül 2021 tarihinde gerçekleştirildi. Atıf indeksleri ile ilgili olarak Social Sciences Citation Index, Science Citation Index Expanded ve Emerging Sources Citation Index kullanıldı. Basit arama alanında konu bölümüne "Time Series Regression", "Air Pollution", "Mortality" terimleri "and" kullanılarak yazılıp arandı. Arama 2011-2021 tarihleri seçilerek yapıldı. Ulaşılan makalelerin yayın yılı, başyazarı, yayın tipi, yayınlandığı ülke ve yayın metodunda kullanılan zaman serisi regresyon modeli bilgileri alınıp Excel dosyasına aktarıldı. Elde edilen yayınlarda WOS üzerinden 100 atıf ve üzeri alan yayınlar dikkate alındı. Daha sonra işaretlenen yayınların Altmetrik İlgi Skorları (AİS) (Altmetric Attention Scores" (AASs)) ve yayın ile ilgili atılan Tweet sayıları, Mendeley platformunda ilgi düzeyleri incelenerek WOS üzerinden aldığı atıf sayıları ile karşılaştırıldı. Altmetrik İlgi Skoru puanlarına "Altmetric Explorer web site" (https://www.altmetric.com/products/explorer-for-publishers/) sitesinden ulaşıldı

Bulgular: Belirtilen anahtar kelimeler girildiğinde son 10 yıl içerisinde WOS üzerinden 539 makaleye ulaşıldı. 100 ve üzeri atıf alan makale sayısı 20 olarak tespit edildi. Atıf sayılarının ortalamasının 143.90±55.82, AİS ortalamalarının 11.00±26.30, yayınlar hakkında atılan tweet ortalamasının 7.95±27.65, Mendeley platformunda okunma ortalamasının 130.85±55.14, yayın üzerinden geçen sürenin 7.60±1.47 olduğu gözlendi. İlgili makalelerin yayınlandığı dergiler incelendiğinde tümünün Q1 kategorisinde dergiler olduğu ve etki faktörlerinin ortalamalarının 8.00±5.67 olduğu görüldü. Yayınlara yapılan atıf sayıları ile AİS arasında pozitif yönde orta düzeyde istatistiksel olarak anlamlı bir korelasyona rastlandı (r=0.491;p=0.028). Benzer olarak AİS ile yayınlara atılan Tweet sayıları arasında da pozitif yönde güçlü anlamlı korelasyona rastlandı (r=0.706;p=0.001). Kullanılan metotlar incelendiğinde yarısına yakınında (%45) GLM-Poisson Regresyon, dörtte bir oranında DLNM ve beşte birinin ise GAM olduğu görüldü.

Sonuç: Son yıllarda yaşanan aşırı iklim olayları, insan hayatının birçok alanını olumsuz yönde etkilemektedir. Hava olayları ile hastalıkların ilişkisini araştırmada kullanılan zaman serisi modelleri ile yapılan makalelere akademinin ilgisinin olduğu kadar sosyal meydanında ilgisinin olduğu görüldü. GLM-Poisson regresyon modellerinin en çok kullanılan modeller olduğu gözlendi.

Anahtar Sözcükler: Çevre epidemiyolojisi, Zaman serisi modelleri, Altmetrik İlgi Skoru

S33

Tip2 Diyabet Hastaları için Yapay Zeka Tabanlı Kişiselleştirilmiş bir Zeki Sistem Önerisi

Osman Tolga KASKATI[1], Burcu TAŞKAN DEMİRKÖK[2], Onur YEŞİL[2], Şeref SAĞIROĞLU[3]

[1]Lokman Hekim Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara
[2]BYS Grup Bilişim Sistemleri Danışmanlık Ticaret ve Sanayi A.Ş., Ankara
[3]Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara

Özet

Giriş: Tip 2 Diyabet Mellitus (T2DM); toplumda 40 yaşın üzerindeki kişilerde sıklıkla görülen özellikle de ülkemizde hızla artan diyabet türlerindendir. Pankreasın yeterli miktarda insülin salgılayamaması veya salgılanan insülinin yeterli derecede kullanılmaması nedeniyle kan şekerinin yükselmesi ile sinir sistemi, göz, kalp ve kan damarları gibi uzuvlarda sıkıntılara veya kayıplara yol açabilmektedir. Dünya'da 500 milyona yakın diyabet hastasının bulunduğu gözönüne alındığında; T2DM gibi kronik hastalıklar, tedavi süreçlerinde ilaçların yanı sıra kişinin yaşam kalitesinin yükseltilmesi, bunu yükseltmede bilişim teknolojileri (BT) tabanlı çözümler geliştirme, yaşam faaliyetlerimizi takip ederek toplanan verileri yapay zeka (AI) temelli modeller ile analiz edilebildiği ve sonuçta bireye özgü teknolojik çözümler ile yaşam kalitesinin iyileştirilmesine yönelik çözümlere ihtiyaç vardır. Bu çalışmada, AI tabanlı modelleme yaklaşımları ile T2DM hastalığının iyileştirilmesine yönelik bireye özgü zeki bir çözüm geliştirilmiş ve kullanıma sunulmuştur.

Yöntem: Çalışma kapsamında T2DM hastalarının yaşam kalitesini artırmaya yönelik olarak zeki bir sistem geliştirilmiştir. Zeki sistem, T2DM diyabete ilişkin bireye özgü sorular sorup, cevap alabileceği yapay zeka tabanlı bir sohbet robotu (chatbot) olup, hastaların gerek uyarılmasında gerekse ihtiyaç duyulan anda durumuyla ilgili cevap alabileceği ve sonuçta davranışlarını kontrol edebileceği bir yapı olarak tasarlanmıştır. Sohbet robotu, hastalardan toplanan veriler ile doktorların verdiği kararlardan oluşan veriler dikkate alınarak eğitilen çok katmanlı yapay sinir ağı ve doğal dil işleme (NLP) yöntemi kullanılarak sorular anlamlandırılmakta ve ayrıştırılmakta, soruların tanımlanması sonrası belirlenen kurala bağlı olarak analiz edilmekte ve elde edilen cevaplar hastaya anında iletilmektedir. Hastanın genel durumu ve doktor tavsiyesi göz önünde bulundurularak periyodik aralıklarla hastaya karar bildirimleri de iletilmektedir. Bu şekilde, kullanıcının yapmak istediği eylemi diyalog kurarak yapan, kullanıcılarına insanlarla konuşuyormuş gibi deneyim yaşatan bir altyapı oluşturulmuştur. Sohbet robotu; çapraz platform (cross-platform) mobil platformu olacak şekilde hem "IOS" ve "Android" programlama dili hem de "React-Native" programlama çerçevesi olarak 3 dilde hazırlanmış olup arka yüzde (back-end) mikro servis yapısında Python Flask microframework tercih edilmiştir. Arka yüz tarafında, AI tabanlı soru tanıma, soruların tanımlanması sonrası belirlenen kurala bağlı kapsam devamlılığını sağlayan veri tabanı modülü ve kullanıcıya hatırlatma için kullanılan sohbet robotu bildirim sistemi bulunmaktadır. Geliştirilen zeki sistem ile, NLP yöntemleri ile veri önişleme süreçlerinden sonra veri tabanı oluşturulur, yapay zeka ile sorulan sorunun karşılaştırılması yapılır ve bağlam çıkarılır; soru havuzu genişletilebilir bir yapıda hazırlanmış olup yeni veriler sisteme eklendikçe de model tekrarlanır.

Bulgular ve Sonuç: Veri önişleme bünyesinde 109 tane eşsiz kelime kökü elde edilmiştir. Eğitilen model sonucunda kullanıcının sorduğu soru belirlenen 23 bağlam için olma olasılığı hesaplanmıştır. Model testleri ve doğrulaması sonucunda elde edilen başarı oranı 50 soruluk test bünyesinde %93 olarak ölçülmüş olup çözümün; hastaların yaşam kalitesinin artırmanın yanında, sağlık profesyonelleri ve bakım merkez çalışanlarının işlerini kolaylaştıracağı değerlendirilmektedir. Sunulan çözümün, T2DM hastalarını güçlendirme, diyabetin artan insidans oranları ile mücadele etmeyi kolaylaştırma, komplikasyonları azaltma ve tüketilen kaynakları minimize etme kullanılabileceği ve özellikle de T2DM hastaları için etkili, kişiselleştirilmiş ve otomatize edilmiş yeni bir çalışma olduğu değerlendirilmektedir.

Anahtar Sözcükler: Yapay Zeka, Tip 2 diyabet mellitus, çok katmanlı yapay sinir ağı, doğal dil işleme, sohbet robotu

Kaynaklar:

1. AS Lokman, JM Zain, FS Komputer, K Perisian: "Designing a Chatbot for Diabetic Patients". International Conference on Software Engineering & Computer Systems, 2009

2. AS Lokman, MA Ameedeen: "Modern Chatbot Systems: A Technical Review". Proceedings of the future technologies conference, 1012-1023, 2018
3. Prathamesh Kandpal, Kapil Jasnani, Ritesh Raut, Siddharth Bhorge: "Contextual Chatbot for Healthcare Purposes (using Deep Learning) ". In Proceedings of IEEE, 2020
4. Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Shreya Bisen, Vasundhara Rathod, "Implementation of a Chatbot System using AI and NLP". International Journal of Innovative Research in Computer Science & Technology (IJIRCST), May 2018
5. Tobias Gentner; Timon Neitzel; Jacob Schulze; Ricardo Buettner: "A Systematic Literature Review of Medical Chatbot Research from a Behavior Change Perspective". 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), July 2020
6. Yuriy Gapanyuk, Sergey Chernobrovkin, Aleksey Leontiev, Igor Latkin, Marina Belyanova, and Oleg Morozenko, "The Hybrid Chatbot System Combining Q&A and Knowledgebase Approaches". AIST 2018

S34

An Alternate Approach to Meta-Analysis for the Ranking of Risk Factors on the Same Dataset for COVID-19 Time-To-Event Data

Dr. Öğr. Üyesi. Ayşe ÜLGEN

Girne Amerikan Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Karmi, Cyprus
Mail: ayshe.ulgen@global.t-bird. edu, ORCID: 0000-0002-0872-667X
Dr. Öğr. Üyesi Hatice Dörtok Demir
Amasya Universitesi, Tıp Fakültesi, Medikal Biyokimya Anabilim Dalı, Amasya
Mail: hatice.demir@amasya.edu.tr, ORCID: 0000-0002-1864-2410
Doç. Dr. Sirin ÇETİN
Tokat GaziosmanPaşa Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Tokat
Mail: cetinsirin55@gmail.com, ORCID: 0000-0001-9878-2554
Dr. Wentian Li
The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York, Amerika Birleşik Devletleri
Mail:wtli2012@gmail.com, ORCID: 0000-0003-1155-110X

ABSTRACT

Aim: Determining which factors contribute to COVID-19 patients' hazard (shorter time to mortality) is important both for the understanding of the disease, but also hospital management. Most often, the single-variable test p-value is used for ranking risk factors. On the other hand, the important role played by a variable in a single or multi-variable classifier can be another way to rank factors. Furthermore, a factor that is protective in leading to a faster release from the hospital provides an alternative perspective.

Material and Methods: More than 3000 people and 30 potential risk factors were collected from the Amasya State Hospital between March and November 2020. The 2682 outpatients do not have time-to-event data and were excluded from survival analysis. We create a composite ranking for a COVID-19 time-to-event data by averaging ten measures in four different methods: two measures (discordant index and integrated Brier score) in single-variable RSF (random survival forest), single-variable Cox proportional hazard ratio model , permutation based variable importance in the full RSF model, and the order of variable selection in the regularized regression LASSO (least absolute shrinkage and selection operator), applying to time-to-death and time-to-release events respectively. With the composite ranking order, we carried out a forward variable addition (top ranked factors added first) and backward variable reduction (lowest rank factors removed first) in RSF to choose a cutoff of the list.

Results: Using the error curve we roughly estimate the number of relevant factor groups to be from 9 to 17. The top-5 factor groups for event-to-death analyses are: urea/creatine, calcium, neutrophil/white-blood-cell, lymphocyte, and glucose; the top 5 factor groups for event-to-release analyses are: ferritin, red-cell-distribution-width, age, neutrophil/white-blood-cell, and calcium. The overall top-5 factors are calcium, white-blood-cell/neutrophil, urea, d-dimer, and red-cell-distribution-width

Conclusion: We have carried out an analysis on factors affecting COVID-19 hazard or protection in a time-to-event data from a Turkish cohort. We use multiple methods to rank factors; some are traditional single-variable test, such as Cox proportional hazard ratio model, whereas others are multiple-variable machine learning techniques, such as RSF. A novel choice in our composite ranking is to find protective ability of lower-value of a risk factor that causes a faster hospital release. This provides a complement to other approaches that only focus on faster time to death. All of our top choices in the composite ranking list were already discovered by other studies to be associated with COVID-19 severity or/and mortality.

Keywords: Random Survival Forest, COVID-19, Time-to-Event-Data, Survival Analysis, Meta-Analysis

S35

Use of Classification Algorithms as a Support in Clinical Decision Making: An Application on Hypothyroidism-Related Headache

Mehmet Ali SUNGUR[1], Demet Hanife SUNGUR[2], Derya ULUDÜZ[3], Rabia Gökçen GÖZÜBATIK ÇELİK[4], Esra HATİPOĞLU[5]

[1]*Düzce University, Department of Biostatistics and Medical Informatics, Düzce*
[2]*Düzce University, Department of Electrical, Electronics and Computer Engineering, Düzce*
[3]*İstanbul University, Department of Neurology, İstanbul*
[4]*Bakırköy Mental Health and Neurological Diseases Training and Research Hospital, Department of Neurology, İstanbul*
[5]*İstanbul University, Department of Internal Disease, İstanbul*

Aim: Headache is a common disease in the community and is divided into subtypes of primary headache such as migraine and tension-type headache. Headaches seen in patients with Hashimoto's thyroiditis may be a primary headache or hypothyroidism-related headache. The aim of this study is to examine the usability of machine learning approaches in the detection of hypothyroidism-related headaches in patients with Hashimoto's thyroiditis and to compare the performance of classification methods.

Methods: The data of 155 patients diagnosed with Hashimoto's thyroiditis, who followed up in the Endocrinology outpatient clinic of Istanbul University, were analyzed in the study, and a headache was accompanied in 95 patients. Forty-eight of these patients were found to have hypothyroidism-related headaches. The data set consists of age, gender, antibodies, thyroid function tests, disease duration, headache frequency, severity, character, and localization. Among the classification methods, decision trees, random forest, k-nearest neighbor, naive Bayes, and support vector machines were used. The cross-validation method was applied to prevent the over-learning problem in the data set. Accurate classification rate, accuracy, sensitivity, and selectivity criteria were used to evaluate the classification performance of the models used in the study.

Results: While 0.8632 accuracy was obtained with the random forest algorithm, an accuracy of 0.8105 with the k-nearest neighbor algorithm; an accuracy of 0.8842 with the naive Bayes algorithm, and an accuracy of 0.9263 with the support vector machines were obtained. Sensitivity and selectivity were obtained as 0.8542 and 0.8723 with the random forest algorithm; 0.7708 and 0.8511 with the k-nearest neighbor algorithm; 0.8333 and 0.9362 with the naive Bayes algorithm; and 0.9375 and 0.9149 with the support vector machines, respectively.

Conclusion: It has been observed that the support vector machines perform better classification compared to other algorithms. The classification successes obtained in terms of using classification algorithms in clinical decision support systems seem promising.

Keywords: Machine learning, classification, hypothyroidism-related headache.

S36

Gen İfade Veri Setlerinde Boyut İndirgeme Yöntemlerinin Sınıflama Performansına Etkilerinin Karşılaştırılması

Fatma Hilal YAĞIN[1], Harika Gözde GÖZÜKARA BAĞ[1]

[1]*İnönü Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya, Türkiye.*

Giriş: Bu çalışmanın amacı, yüksek boyutlu Akut Miyeloid Lösemi (AML) hastalığı gen ifade veri setinde boyut indirgeme yöntemlerinin (temel bileşenler analizi (PCA) ve bağımsız bileşenler analizi (ICA)) çeşitli destek vektör makinesi sınıflandırma yöntemlerine etkilerinin karşılaştırılmasıdır.

Yöntem: Bu çalışmada GEO veri deposunda GDS3057 kodu ile yüklenen AML gen ifade veri seti kullanılmıştır. Veri setinde 38 sağlıklı donörden alınan normal hematopoietik hücreler ile 26 AML hastasından gelen lösemik blastlar arasındaki gen ifade profilleri bulunmaktadır. AML veri seti 64 kişi ve 22283 gene ait ifade seviyelerini içermektedir. Veri setine filtreleme işlemi yapıldıktan sonra temel bileşenler analizi (PCA) ve bağımsız bileşenler analizi (ICA) yöntemleri uygulanmıştır. Bu yöntemlerden elde edilen boyutu indirgenmiş veri setleri kullanılarak Doğrusal, Polinomiyal ve Radyal tabanlı çekirdek fonksiyonlu Destek Vektör Makinesi (DVM) yöntemleri ile üç farklı model oluşturulmuştur. Modelleme aşamasında yeniden örnekleme yöntemi olarak 10 tekrarlı 10 katlı çapraz geçerlik ve hiperparametre optimizasyonu için rasgele arama yöntemi kullanılmıştır. Oluşturulan modellerin performansını değerlendirmek için doğru sınıflama oranı, duyarlılık, seçicilik, kesinlik ve F1 skor değerlerinin ortalamaları verilmiştir.

Bulgular: Filtreleme işlemi yapıldıktan sonra AML veri setinde 6201 gen kalmıştır. PCA/ICA uygulandıktan sonra AML gen ifade veri setinden 10 bileşen çıkarılmıştır. Sonuçlar incelendiğinde AML hastalığını ayırt etme gücü en iyi olan model veri setine PCA uygulandıktan sonra Polinomiyal çekirdek fonksiyon ile kurulan DVM modeli olarak bulunmuştur (Doğruluk: % 95.64, Duyarlılık: % 94.53, Seçicilik: % 96.46, Kesinlik: % 95.48, F1 Skor: % 94.99).

Sonuç: Gen ifade veri setleri ile sınıflandırma modelleri oluşturulmadan önce boyut indirgeme yöntemleri kullanılarak yüksek boyutluluk sorunu giderilmeli, modeller daha sonra oluşturulmalıdır. Bu yöntemler analiz süresini kısaltır ve modellerin ayırt etme gücünü arttırır. Bu çalışmada; AML gen ifade veri setinde sınıflandırma modellerinden Polinomiyal çekirdek fonksiyon ile kurulan DVM modelleri, Doğrusal ve Radyal tabanlı çekirdek fonksiyonu ile kurulan DVM modellerine göre daha iyi sonuç vermiştir. Ek olarak PCA'nın ICA'ya kıyasla model performansına katkısının daha fazla olduğu söylenebilir. Ancak birden fazla veri setinde ve/veya simüle veri setinde bu yöntemleri deneyerek sonuçları karşılaştırmak daha kesin sonuçlara ulaşılması açısından önemlidir.

Anahtar Sözcükler: Boyut İndirgeme, Gen İfade Veri Seti, Özellik Çıkarımı, Sınıflandırma

KISA SÖZLÜ BİLDİRİLER
SHORT ORAL PRESENTATIONS

KS1

A Nomogram Model to Predict Bone Metastasis in Patients with Breast Cancer

Senem KOÇ[1], Halil Ibrahim TANBOGA[1], Fuzuli TUGRUL[2]

[1]Nisantasi University, Faculty of Medicine, Department of Medical Statistics, İstanbul
[2]Eskisehir City Hospital, Department of Radiation Oncology, Eskisehir

Introduction: Breast cancer is the most common cancer in women worldwide. These patients present with metastatic disease at the time of first diagnosis, with early-stage disease will afterward develop metastasis. Breast cancer often metastasizes to the bone. The aim of the study was to use clinical variables to develop a logistic regression nomogram that can predict bone metastasis with breast cancer.

Methods: Medical records of patients with breast cancer were retrospectively collected between 2018 and 2021 in Eskisehir City Hospital. A Logistic regression nomogram was performed using R.

Results: Among 210 patients with breast cancer, 47 developed bone metastases. The results showed that the major risk factors of bone metastases (Fig. 1) are Age, Ki67, oestrogen, progesterone, Cerb2, monocytes, lymph node status, and menstruation status which were significantly and independently associated, the area under ROC was 0.716.

Conclusions: We have developed a clinical nomogram to predict bone metastatic breast cancer and this model can be printed out and used for probability prediction without the use of computer or calculator.

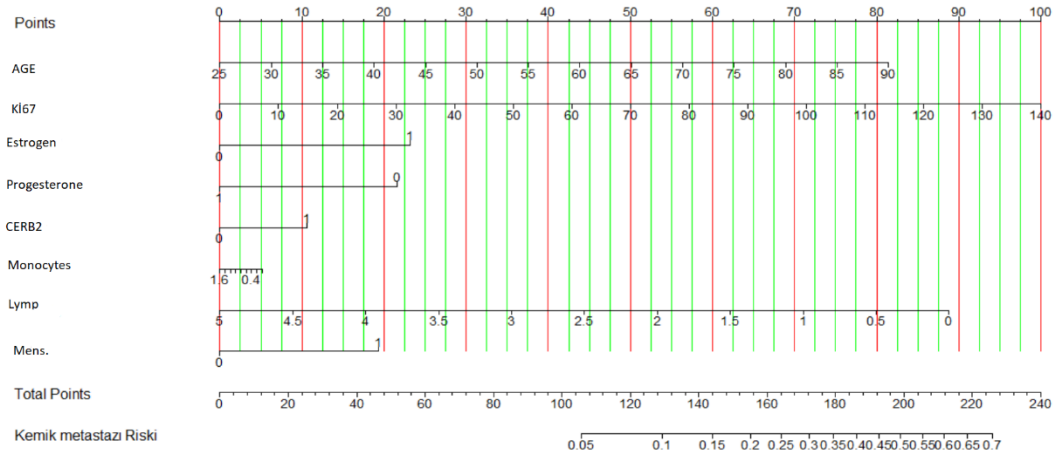


Fig.1. Nomogram to predict the probability of Bone Metastasis in Patients with Breast Cancer

Keywords: Breast cancer, bone metastasis, nomogram, logistic regression, prediction.

References:

1. Y. C. Chen, D. M. Sosnoski, A. M. M. Chen. Breast cancer metastasis to the bone: mechanisms of bone loss. Breast Cancer Research, 2010, 12:215.
2. G. Tinoco1, S. Warsch, S. Glück, K. Avancha, A. J. Montero. Treating Breast Cancer in the 21st Century: Emerging Biological Therapies. Journal of Cancer, 2013; 4(2): 117-132.
3. N. Howlader, A. M. Noone, M. Krapcho. SEER Cancer Statistics Review, 1975–2018. National Cancer Institute. 2021.
4. M. Xue, Y. Su, Z. Feng, S. Wang, M. Zhang, K. Wang, H. Yao. A nomogram model for screening the risk of diabetes in a large-scale Chinese population: an observational study from 345,718 participants. Scientific reports, 2020, 10:11600.

5. A. Zlotnik, V. Abraira. A general-purpose nomogram generator for predictive logistic regression models. The Stata Journal, 2015, 15 (2), 537–546.
6. S. Pu, K. Wang, Y. Liu, X. Liao, H. Chen, J. He, J. Zhang. Nomogram-derived prediction of pathologic complete response (pCR) in breast cancer patients treated with neoadjuvant chemotherapy (NCT). Pu et al. BMC Cancer (2020) 20:1120.

KS2

Investigation of the Effect of Hepatosteatosis on Severity Score Obtained by Repeated Measures in COVID-19 Patients

Şirin Çetin[1], Özge Pasin[2], Ahmet Turan Kaya[3]

[1] Biostatistics Department, Faculty of Medicine, Tokat GaziosmanPasa University, cetinsirin55@gmail.com
[2] Biostatistics Department, Faculty of Medicine, Bezmialem University, opasin@bezmialem.edu.tr
[3] Radiology Department, Amasya State Hospital, hmttrnky.62@gmail.com

Abstract

Introduction

The aim of this study is to examine the change in the computed tomography (CT) score value of patients with COVID 19 and hepatosteatosis (nonalcoholic fatty liver disease). For this purpose, outpatients and inpatients from Amasya State Hospital were followed, and COVID-19 patients whose diagnosis was confirmed by laboratory were retrospectively screened. Patients' demographic information, comorbidities, history of hospitalization or intensive care unit (ICU) admission, laboratory findings, length of hospital stay, and electronic medical records were recorded. According to electronic medical records, laboratory examinations obtained within 1 day of the date corresponding to the first chest CT scan at the time of admission to the hospital. The first Total CT Severity Score values were obtained from patients on recurrent days. In some patients, 1 score was obtained from some patients, 2 from some patients, and 3 from some patients. It is desired to investigate whether these scores change the scores in patients with and without hepatosteatosis.

Materials and Methods

For this reason, mixed analysis was used in the analysis. In the data obtained from repeated score measurements, as a result of the inability to obtain 2nd and 3rd score values from some patients, missing (lost) valuable data occurred. If we do this analysis with the GLM method, 348 patients with missing values out of 440 will be excluded from the analysis and the analysis will continue with 92 patients. In this case, many data will be excluded from the analysis. To deal with this data. In other words, MIXED model analysis was used to include patients with missing data in the analysis. The mean age of our patients in the study was found to be 60±12.08. 60% of the patients are male and 40% are female. 10% of the patients were treated as outpatients, 90% were treated as inpatients.

CT score value was found as 16 threshold by ROC analysis. The hospital stay was found to be 26 (95% CI, 21.42-30.33) in patients with a CT score of ≥ 16, and 22 days (95% CI, 17.43-26.75) in patients with a CT score of < 16. The length of stay in hospital for patients with hepatosteatosis was 29(95% CI, 23-34) The length of stay in hospital for patients without hepatosteatosis was 24((95% CI-19-28) (patients with CT score ≥16 compared to patients with CT score <16) It is approximately 2 times more mortal, and those with hepatosteatosis are approximately 2 times more mortal than those without.

Discussion and Conclusions

As a result of the MIXED analysis performed to see the change of hepatosteatosis on the CT score value, we can state that hepatosteatosis is significant on the CT score (p=0.0052) and this change increases over time in the 2nd and 3rd score values. As a result, we can say that the CT scores of patients with hepatosteatosis with COVID 19 started to increase rapidly in the following days, especially after the 7th day.

Keywords: Repeated Measures, COVID 19, Hepatosteatosis, CT Severity Score Values

References

1. Littell, R. C., Henry, P. R., & Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. Journal of animal science, 76(4), 1216-1231.
2. Littell, R. C. (1990). Analysis of repeated measures data.
3. Albert, P. S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. Statistics in medicine, 18(13), 1707-1732.

4. Babalola, O. E., Bode, C. O., Ajayi, A. A., Alakaloko, F. M., Akase, I. E., Otrofanowei, E., ... & Omilabu, S. A. (2021). Ivermectin shows clinical benefits in mild to moderate Covid19 disease: A randomised controlled double blind dose response study in Lagos. medRxiv.
5. López-Medina, E., López, P., Hurtado, I. C., Dávalos, D. M., Ramirez, O., Martínez, E., & Caicedo, I. (2021). Effect of ivermectin on time to resolution of symptoms among adults with mild COVID-19: a randomized clinical trial. Jama, 325(14), 1426-1435.
6. Moonis G, Filippi CG, Kirsch CFE, Mohan S, Stein EG, Hirsch JA, et al. The spectrum of neuroimaging findings on CT and MRI in adults with COVID-19. Am J Roentgenol. 2021;217(4):959–74.
7. Mahajan A, Hirsch JA. Novel coronavirus: what neuroradiologists as citizens of the world need to know. Am Soc Neuroradiology; 2020.
8. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3.
9. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science (80-). 2020;367(6485):1444–8.

KS3

A Spatial Analysis of Air Quality and Food Consumption Surrogates in Relation to Prostate Cancer Deaths

Selman Aktas[1,2], **Mehmet Kocak[2,3,*]**

[1] *Ph.D. Candidate Department of Biostatistics, School of Medicine, Istanbul Cerrahpasa University*

[2] *Biostatistics And Medical İnformatics,Health Sciences University, Istanbul, Turkey*

[3] *Ph.D. Professor of Biostatistics,Division of Biostatistics and Medical Informatics, International School of Medicine, Istanbul Medipol University*

[*] *Corresponding Author: mehmetkocak@medipol.edu.tr*

http://orcid.org/0000-0002-3386-1734

DECLARATIONS

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

Conflicts of interest/Competing interests: The authors have no relevant conflict of interest.

Availability of data and material: As the death records and fruit and vegetable consumption data used in this report were granted access only to the corresponding author, we do not have permission to share these data components; however, we can share the air quality data upon request.

Code availability: Not applicable

Ethics approval: Not applicable as only publicly available data were used

Consent to participate: Not applicable as this study has no direct patient data and only publicly available data is used

ACKNOWLEDGEMENT

This study was partially funded by TUBITAK Directorate of Science Fellowships and Grant Programmes (BİDEB)-2232 International Fellowship for Outstanding Researchers. We also thank the Turkish Republic Ministry of Commerce and Turkish Statistical Institute for data sharing. The opinions raised in this article solely belong to its authors, and does not represent the position of TUBITAK, Turkish Republic Ministry of Commerce and Turkish Statistical Institute in any shape or form.

Abstract

Objective: Objective: American Cancer Society reports that about 1 in every 8 men experiences prostate cancer diagnosis in his lifetime and 1 in every 41 men dies of prostate cancer on average. Common risk factors for prostate cancer are age, race, geographic location, and genetic factors. The causality of other factors such as dietary habits, overweight and obesity, smoking and chemical exposure are not clear and therefore, there is still unmet knowledge gap requiring further investigations in a spatial and temporal manner. We carried out such a study to explore the association of meteorological and air-quality markers, fruit and vegetable consumption and drinking water source trajectories with prostate cancer deaths in spatial models. **Material and Method:** Province-level data on prostate cancer deaths, meteorological markers such as such as air-pressure, humidity, temperature, etc, air-quality markers, namely, particular matter 10 and 2.5 (PM_{10}, $PM_{2.5}$), sulfur-dioxide (SO_2), carbon-monoxide (CO), nitrogen-dioxide (NO_2), and ozone (O_3), drinking water source data from rivers, dams, wells, and springs as well as fruits and vegetables sales data were obtained for 81 provinces of Turkey for Years 2018 and 2019. Mixed modelling approach taking into consideration the spatial autocorrelation was used to investigate the associations of prostate cancer deaths with these environmental and food consumption variables. **Results:** These models revealed increased prostate cancer deaths with increased humidity, number of rainy days, and mean temperature. Similarly, provinces with higher concentration of CO, NO_2 and O_3 reported higher Prostate Cancer deaths, while provinces

having higher values of PM_{10} reported lower number of prostate cancer deaths. Finally, Provinces with higher consumption of beans, dill, grape, lemon, purslane, and strawberry reported lower prostate cancer deaths. When considered in a multivariable model, only humidity retained its significance. **Conclusion:** As a conclusion, although several air-quality and dietary consumption variables showed univariable significance in relation with prostate cancer deaths, only average humidity in a given locality seems to be significantly associated with prostate cancer deaths.

Keywords: prostate cancer deaths, meteorological markers, air quality markers, drinking water sources, fruit consumption, vegetable consumption

Introduction

American Cancer Society reports that about 1 in every 8 men experiences prostate cancer diagnosis in his lifetime and 1 in every 41 men dies of prostate cancer on average. Globally, prostate cancer is the most commonly diagnosed cancer among men with approximately 1.6 million cases in 2015[1]. International Agency for Research on Cancer reported in 2018 that "with an estimated 1,276 000 new cancer cases and 359 000 deaths in 2018, it [prostate cancer] is the second most frequently diagnosed cancer among men and the sixth leading cause of cancer death among men worldwide."[2] The institute also projects that such figures will double by year 2040 with new prostate cancer diagnosis of 2,300,000 new cases and 740,000 prostate cancer deaths, solely due to growing and aging world population.[2] Prostate cancer is especially common in developed countries partially due to advanced diagnosis techniques being more readily available. Prostate cancer screening is almost routine in developed countries which leads to higher reported incidences.[3] Luo et al. look at the association between socio-demographic index (SDI) and prostate cancer incidence and deaths in spatial models and report that there exist significant spatial autocorrelation for the incidence of and deaths due to prostate cancer, and the incidence is significantly higher in countries with higher SDI while the deaths are significantly lower.[4] This means that higher SDI counties detect prostate cancer earlier and provide better management for its treatment.

Along the same lines, the prostate cancer rate in Western countries is 10-15 times higher than in Asian countries, and most of the prostate cancer-related deaths occur in developed countries; however, in the last decade, Prostate cancer and related mortality in Southeast Asia has increased as this region gradually adopts a western lifestyle, including sedentary habits and a high-fat diet [5]. Advanced age, black race, and family history are well-known risk factors for prostate cancer [6]. In addition, various environmental risk factors such as diet, lifestyle, obesity, metabolic syndrome, sexual behaviors and infections have been reported as associated with prostate cancer [7,8].

While research in the literature regarding the impact of environmental factors on prostate cancer etiology is expanding, there is still a large knowledge gap in the spatial and temporal investigation of environmental factors. A part of the issue is the lack of granular data on the time (temporality) and location (spatiality) domain. As nicely illustrated by Luo et al., spatial analysis approach has started to be used frequently in studies where parameters or variables are thought to be geographically related to each other.[4] These spatial models aim to explain the interaction, structure and processes of the spatial data and to identify their possible relations with other spatial events.[9] Spatial analysis is inspired by the theory that "everything is related to everything else, things that are close are more related than things that are far away" [10].

To help meet this unmet knowledge gap, the primary objective of this paper is to explore the potential associations between various environmental and dietary factors and deaths due to prostate cancer in Turkey in a spatiotemporal manner.

Materials and Methods

The primary outcome variable of this study is the provincial death rates due to prostate cancer. The death rate data were provided for Years 2018 and 2019 by the Turkish Statistics Institute. The environmental and food consumption dimensions of our data include four main sets of variables requested from the Ministry of Agriculture and Forestry, General Directorate of Meteorological Services: Meteorologic markers (air pressure, humidity), predicted and manual counts of rainy days, temperature (maximum, average, and minimum), and wind speed. The next set of variables are the air-quality markers including Particular Matter 10 and 2.5 (PM_{10} and $PM_{2.5}$), sulfur dioxide (SO_2), Carbon Monoxide (CO), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2) as well as NO_X, and Ozone (O_3), and they were manually extracted from the periodic Environmental Impact Evaluation (EIE) reports

generated by the Turkish Ministry of Environment and Urbanization for each of the 81 provinces of Turkey for years 2017-2019. The next dimension of our predictors included local fruit and vegetable sales in local markets within each of the 81 provinces covering 45 different fruits or vegetables, provided by the Turkish Ministry of Commerce for Years 2018-2020. The last set of variables included the amount of drinking water per capita coming from streams, dams, lakes, ponds, springs, and wells. This data was also manually extracted for each of the 81 provinces from the periodic Environmental Impact Evaluation (EIE) reports generated by the Turkish Ministry of Environment and Urbanization for each state on an annual basis between years 2000-2018, and provinces were categorized based on usage trajectories over the years.

The provincial Prostate cancer's data and predictors from the four dimensions mentioned above were merged by province and year identifiers. Food consumption and drinking water source markers were normalized by the population size of each province for each year appropriately leading to per capita consumption expression such as 'average consumption in kg per person per year'. The air-quality and meteorologic markers were expressed as annual averages within each province. Representative maps of different data dimensions have been provided in Figure-1 and Figure-2.

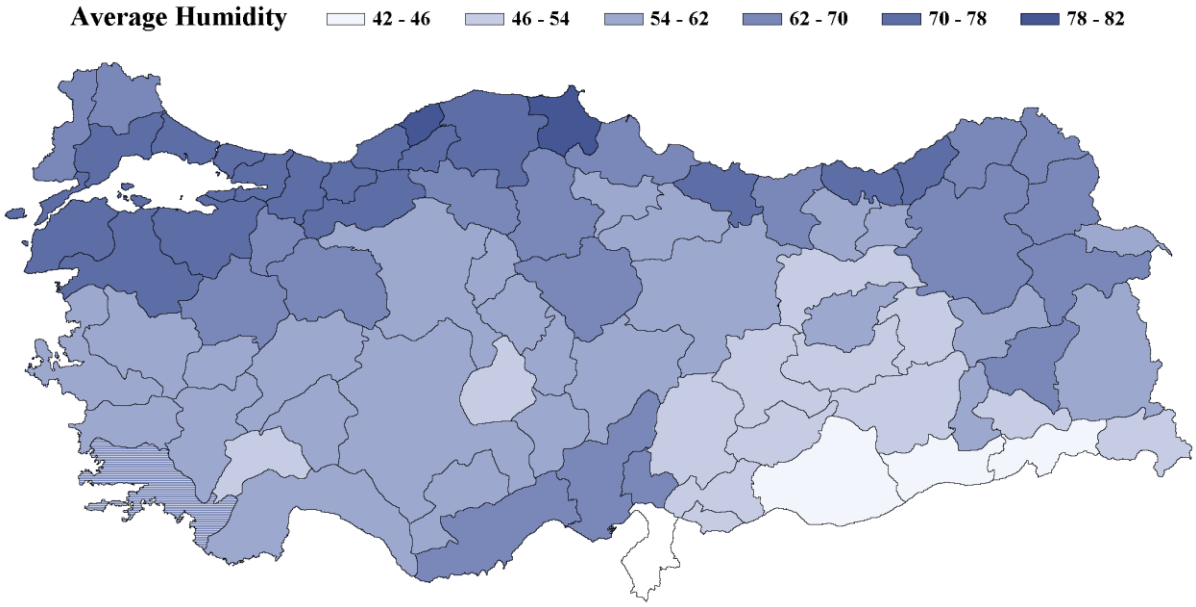


Figure 1. Average Air Humidity

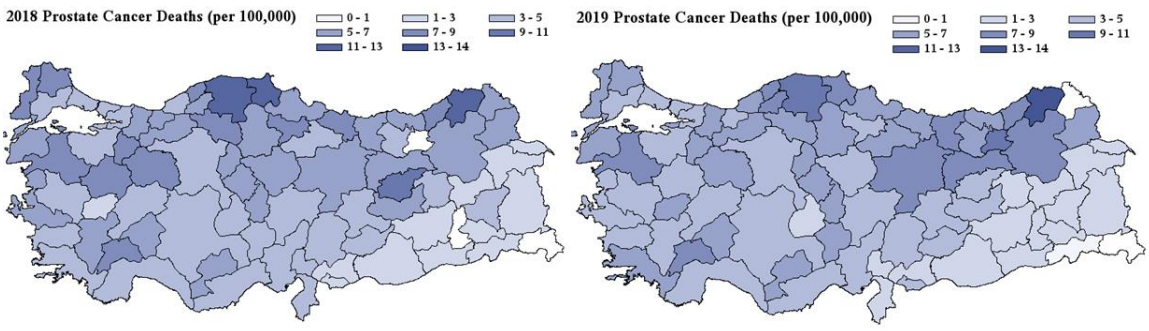


Figure 2. Deaths due to Prostate cancer per 100,000 population for Year-2018 and Year-2019.

As the provincial Prostate cancer deaths per 100,000 population is of spatial nature, we tested for the spatial autocorrelation for each of the 9 timepoints; the outcome data was obtained using Moran's I statistic [11]. For each time point, spatial autocorrelation was highly significant ($p<0.0001$) as seen in Table-1.

Year	Coefficient	Observed	Expected	Std Dev	Z	Pr > |Z|
Year	Moran's I	0.084	-0.013	0.0157	6.18	<.0001
	Geary's c	0.858	1.000	0.0284	-5.00	<.0001
Year	Moran's I	0.111	-0.0128	0.0157	7.94	<.0001
	Geary's c	0.817	1.0000	0.0275	-6.65	<.0001

This significant spatial autocorrelation was controlled for via mixed models with spatial-power covariance SP(LIN) structure using the central latitude and longitude values of each province. We used MIXED procedure in SAS Version 9.4 (®, Cary, North Caroline, USA).

All of the predictors were continuous variables in their original form. However, we have also obtained ordinal versions of these markers by representing them as quartiles (or tertiles); thus, under these definitions, each province fell under one of the four (or three under the tertile definition), categories representing low-exposure or consumption, intermediate-exposure or consumption, or high-exposure or consumption, etc. This approach potentially helped to eliminate the impact of potential outliers in predictors and help with interpretations of the modeling results. An example of such signatures is shown for O_3 in Figure-3.

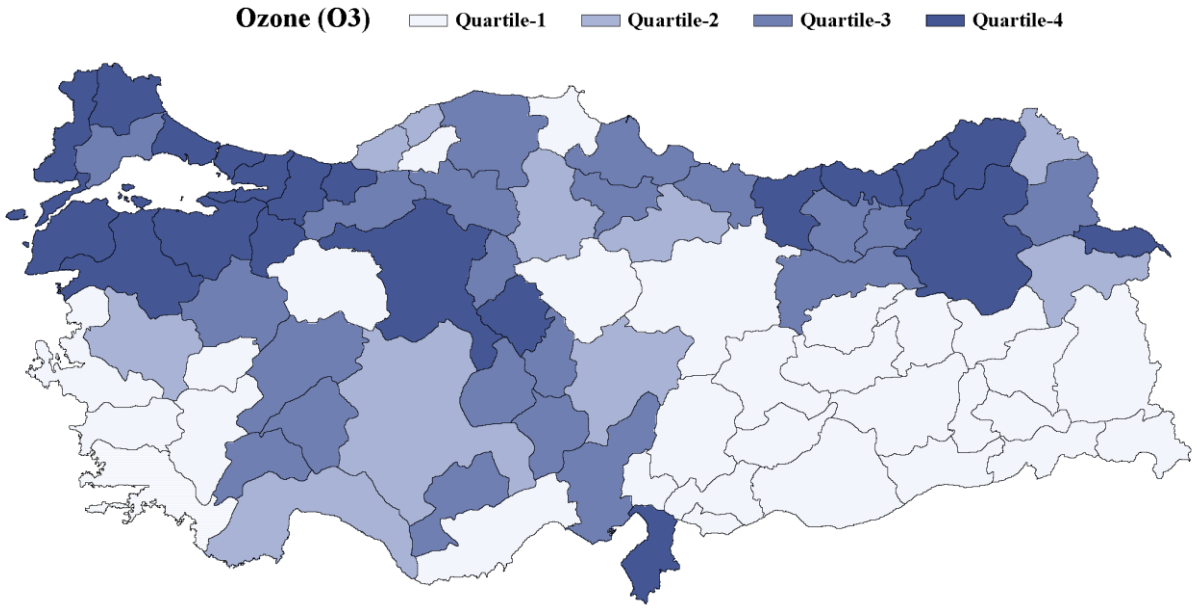


Figure 3. Quartiles of Ozone (O_3)

In addition to the above 'quartile' or 'tertile' approach, whenever there is sufficient longitudinal data, we built trajectories for a given marker using the TRAJ procedure developed for SAS by Jones and Nagin [12]. This approach was feasible for PM_{10}, SO_2, NO_2, O_3 and $PM_{2.5}$ for the air quality markers, for all fruit and vegetable consumption variables using the data from Years 2018-2020, and for the drinking water source, namely, Dam, Spring, Well, and River. We present an example of such trajectories in Figure 4.

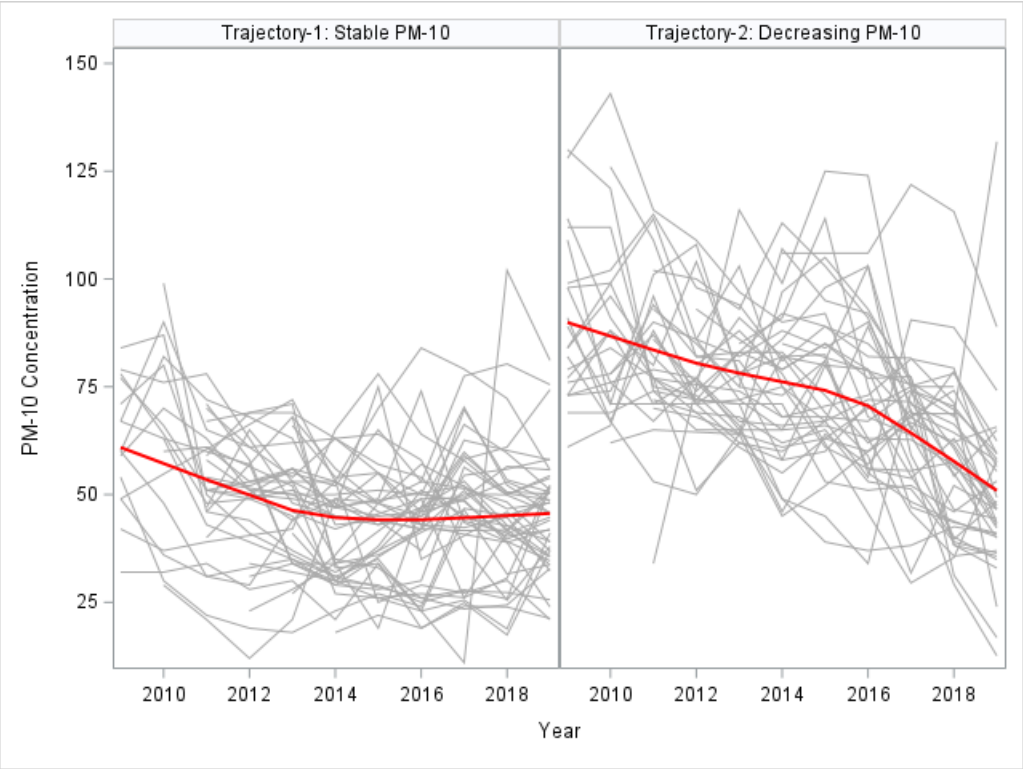


Figure 4. Change trajectories based on annual averages of PM-10 Concentrations

Overall, the list of the predictors included 8 meteorologic markers, 6 air-quality markers, 4 drinking water source profiles, and 45 fruit and vegetable consumption markers including the total fruit and vegetable consumption. Having a large number of predictors, we used the Benjamini-Hochberg False Discovery Rate (FDR) approach [13] to control the familywise error rate within each phase of our model-building approach. To build our models, we started our variable selection approach with the meteorologic markers first. Once significant meteorological markers were identified in univariable models, we tested for the joint significance of these markers and reduced the final set of significant variables as indicated by these multivariable models. We then expanded these multivariable models to include air-quality markers first, then food consumption markers and drinking-water-source profiles. In identifying significant markers, we looked for consistent significance for both Year-2018 and Year-2019, which were modelled independently.

Results

In the first phase of our variable selection approach, we showed that number of rainy days, air pressure and humidity showed somewhat consistent evidence of association with prostate cancer deaths (Table 2, Figure-5), suggesting that provinces with higher concentrations of these meteorological markers reported relatively higher number of prostate cancer deaths.

Table 2. Univariable model estimates with Meteorological Markers

Predictor	Estimate (2018)	Raw P-value (2018)	FDR-adjusted P-value (2018)	Estimate (2019)	Raw P-value (2019)	FDR-adjusted P-value (2019)
Mean Air Pressure	0.01	0.089	0.14	0.00	0.3	0.34
Mean Humidity	0.13	<0.0001	<0.0001	0.12	<0.0001	<0.0001
Mean No. of Rainy Days (Manual)	0.04	<0.0001	0.001	0.04	<0.0001	0.001
Mean No. of Rainy Days (Model-based)	0.02	0.013	0.034	0.02	0.024	0.047
Maximum Temperature	-0.07	0.47	0.54	-0.13	0.18	0.23
Mean Temperature	-0.16	0.052	0.10	-0.23	0.006	0.015
Minimum Temperature	-0.07	0.21	0.28	-0.10	0.09	0.14
Mean Windspeed	0.08	0.86	0.86	-0.11	0.81	0.81

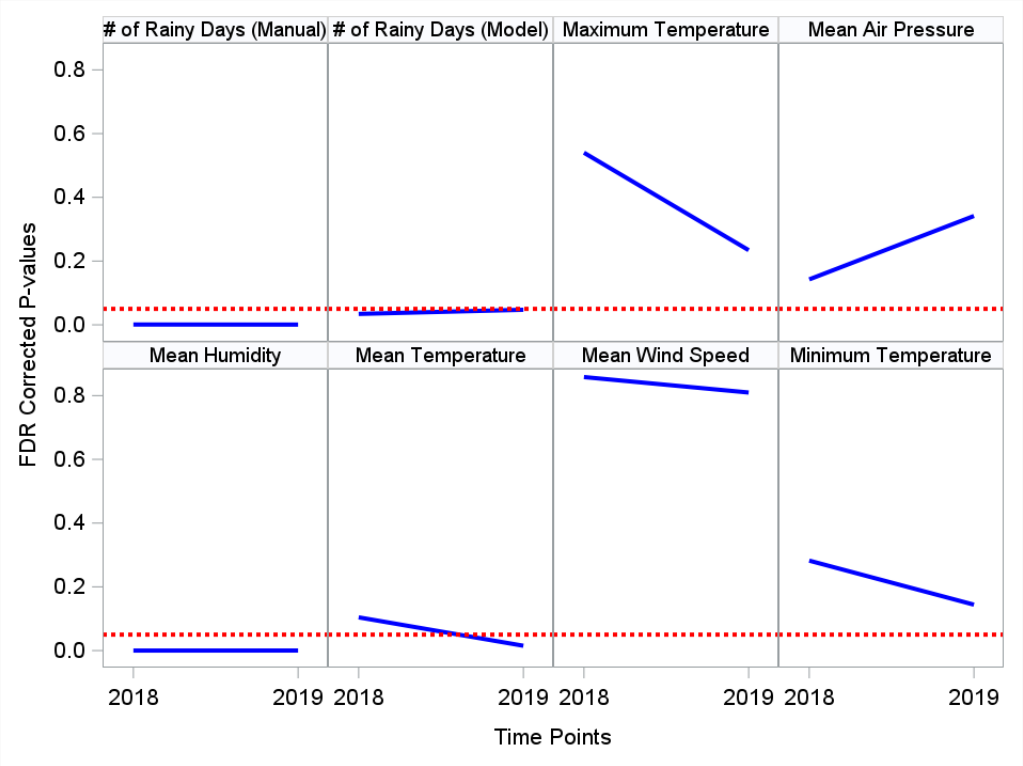


Figure 5. Distribution of FDR-corrected P-values for Meteorological markers

Close evaluation of these markers revealed that only humidity retained its significance in a multivariable model; therefore, humidity will be kept as a primary predictor in multivariable models air quality and food consumption markers.

CO, NO_2, O_3, and PM_{10} quartiles were also found to be univariably significantly associated with prostate cancer deaths (Table 3, Figure 6); this funding indicates that provinces with higher concentrations of these air-markers experienced higher levels of Prostate cancer's deaths except that this association was reversed counterintuitively for PM_{10}.

Table 3. Univariable model estimates with Air-Quality Markers

Predictor	Estimate (2018)	Raw P-value (2018)	FDR-adjusted P-value (2018)	Estimate (2019)	Raw P-value (2019)	FDR-adjusted P-value (2019)
CO quartiles	0.47	0.022	0.043	0.47	0.02	0.039
NO_2 quartiles	0.71	0.001	0.008	0.8	<0.0001	0.001
O_3 quartiles	0.56	0.009	0.028	0.78	<0.0001	0.001
PM_{10} quartiles	-0.52	0.029	0.043	-0.5	0.033	0.049
$PM_{2,5}$ quartiles	-0.27	0.28	0.33	-0.15	0.56	0.67
SO_2 quartiles	0.14	0.55	0.55	-0.05	0.84	0.84

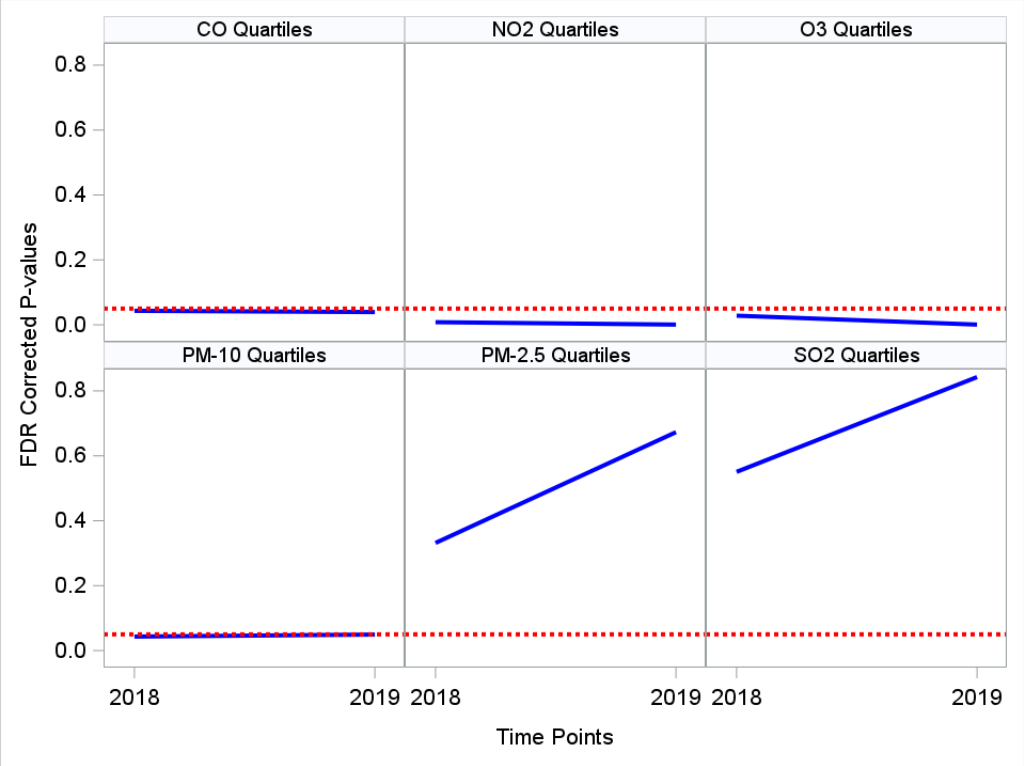


Figure 6. Distribution of FDR-corrected P-values for Air-quality markers

When considered together in the same model with Average Humidity, these markers lost their significance (Supplementary Figure-1).

In univariable spatial models, the consumption of beans, dill, grape, lemon, purslane, and strawberry consumption showed some evidence of significance (Table 4). The provinces that have relatively higher consumption of these fruits and vegetables reported higher number of prostate cancer deaths on average. The rest of the fruit and vegetable markers were presented in Supplementary Figure-2.

Table 4. Food consumption markers showing suggestive significance univariable

	Predictor	Estimate (2018)	Raw P-value (2018)	FDR-adjusted P-value (2018)	Estimate (2019)	Raw P-value (2019)	FDR-adjusted P-value (2019)
Univariable	Beans (kg)	0.70	0.003	0.063	0.80	<0.0001	0.011
	Dill (Bunch)	0.82	<0.0001	0.019	0.80	<0.0001	0.011
	Grape (kg)	0.64	0.005	0.063	0.72	0.001	0.017
	Lemon (kg)	0.59	0.011	0.085	0.69	0.002	0.023
	Purslane (kg)	0.61	0.01	0.085	0.53	0.022	0.095
	Strawberry (kg)	0.66	0.005	0.063	0.75	0.001	0.015
Controlling for Average Humidity	Beans (kg)	0.17	0.49	0.93	0.36	0.16	0.88
	Dill (Bunch)	0.34	0.17	0.93	0.37	0.14	0.88
	Grape (kg)	0.16	0.51	0.93	0.32	0.19	0.88
	Lemon (kg)	0.23	0.31	0.93	0.37	0.10	0.88
	Purslane (kg)	0.22	0.34	0.93	0.14	0.56	0.88
	Strawberry (kg)	0.04	0.88	0.98	0.26	0.33	0.88

Controlling for Humidity, none of these markers remained significant in bivariable models (Lower half of Table 4). Similarly, none of the drinking-water-source markers were found to be consistently significantly associated with Prostate cancer deaths (Supplementary Figure-3).

Discussion

In this study, we have investigated the association of meteorological, air-quality, and fruit and vegetable consumption variables with deaths due to prostate cancer. We have worked with different 'signatures' of these markers, namely, overall province-level crude averages over time, change trajectories, quartiles, tertiles, and median-dichotomization. One immediate advantage of some of these approaches was to reduce the potential impact of outliers in a given predictor on model results, thus leading to more robust and consistent findings over the two years we analyzed, namely Years 2018 and 2019.

Our outcome variable was based on annual counts of prostate cancer deaths in each of the 81 provinces in Turkey. In such spatial data, the first step of analyses is to check the existence of spatial autocorrelation. At this step, we showed that provincial annual prostate cancer deaths per 100,000 population expressed strong spatial autocorrelation among the provinces in both years, which is visible in Figure-2 as provinces with high and low prostate cancer death rates were clustered together on the map. We addressed this spatial autocorrelation using the central longitudes and latitudes of each province (i.e., centroids) in the MIXED procedure in SAS using the REPEATED statement. Contrary to the ordinary least-squares regression models, these models assume correlated errors; that is, spatial autocorrelation is designed to be part of the model errors.

We constructed independent models for Year-2018 and Year-2019 to assess the consistency of the results at least across two consecutive years and reported only the markers that show consistent significance in both years, which represents the indirect temporal component of our analyses. Working with a large number of potential predictors, it is expected to have one spurious finding in every 20 tests on average at Type-1 error rate of 0.05; this necessitates that we recognize and correct this multiplicity issue in our analytic approach; we employed Benjamini-Hochberg

False Discovery Rate (BH-FDR) to control the Familywise Error as much as we can. We also strived to stay away from any claim of causality in our findings and focused more on 'correlation' instead of 'causation'; therefore, our findings must be considered as potential hypotheses to be investigated further in future prospective trials.

One of the main findings from our analyses is the negative association of humidity with deaths due to prostate cancer. Provinces with higher concentration of humidity reported higher numbers of prostate cancer deaths per capita. The only similar study we were able to find was the research by JI and MP [14]. They showed that regions with higher Martonne Aridity Index, which is based on rain fall and temperature, report higher number of prostate cancer deaths. St-Hilaire et al. [15] suggest that "prostate cancer may be partially correlated with meteorological factors. The patterns observed were consistent with what we would expect given the effects of climate on the deposition, absorption, and degradation of persistent organic pollutants including pesticides. Some of these pollutants are known endocrine disruptors and have been associated with prostate cancer."

Controlling for average humidity, none of the other air-quality and food consumption markers remained significance although several markers were univariably significant. As we stayed away from any claim of causality, we aimed to provide suggestive evidence for more targeted prospective studies; therefore, we provided the univariable associations as well. In reality, the significant association we have illustrated for humidity may be due to a totally different mechanism, for example, based on a mixture of temperature and air-quality profile, even food consumptions. Such comprehensive models are needed to tease out the mechanism of combined and synergistic associations with more granular data.

Our study has several weaknesses to mention. Ideally, Prostate cancer diagnosis data would be needed for our intended analyses, but we were able to gain access only to the Prostate cancer deaths data. We were also given access to data only for Years 2018 and 2019, and ideally, longer term data would provide a much stronger statistical power. Beyond the weak time-depth of our data, our data does not have the ideal geographical depth, either, as we have data at provincial level from 81 provinces of Turkey. This limits our sample size to n=81 and thus limits the statistical power of our models. Therefore, in our multivariable approach, we had to be careful in model construction in a step-by-step manner. The second most important weakness of our data is that our food consumption markers are actually local sales data, which are used as surrogate markers for dietary consumption. Eyles et al. [16] showed significant correlation between the market sales data with actual household nutritional intake. Therefore, we still believe that the local market sales data can be reasonable surrogates when individual or household level consumption data are not available. However, we still lack in our data sources in terms of some critical food types such as meat, dairy, and flour-based items. The ideal study design to investigate the association between fruit and vegetable consumption would require granular data including individual level of disease follow-up and individual level of food consumption profiles; however, such data is close to impossible to obtain in a large spatial spectrum like ours and therefore, we were limited by representing each province by a signature of fruit and vegetable consumption using the sales data as surrogates as well as air quality markers with an average behavior over few years. Individual level food-frequency data would be ideal for such analyses. The dietary consumption dimension of our data does not contain many food items such as consumption of other fruits and vegetables, meat, poultry, and fish and their derivatives, as well as dairy products and flour-based products such as bread, pastries, desserts, etc. We were also not able to access provincial water quality data; however, we were able to extract drinking-water source data for each province and used this to represent the water-quality dimension perhaps in a limited manner. We continue working on expanding our data dimensions to increase the depth and breadth of our potential markers.

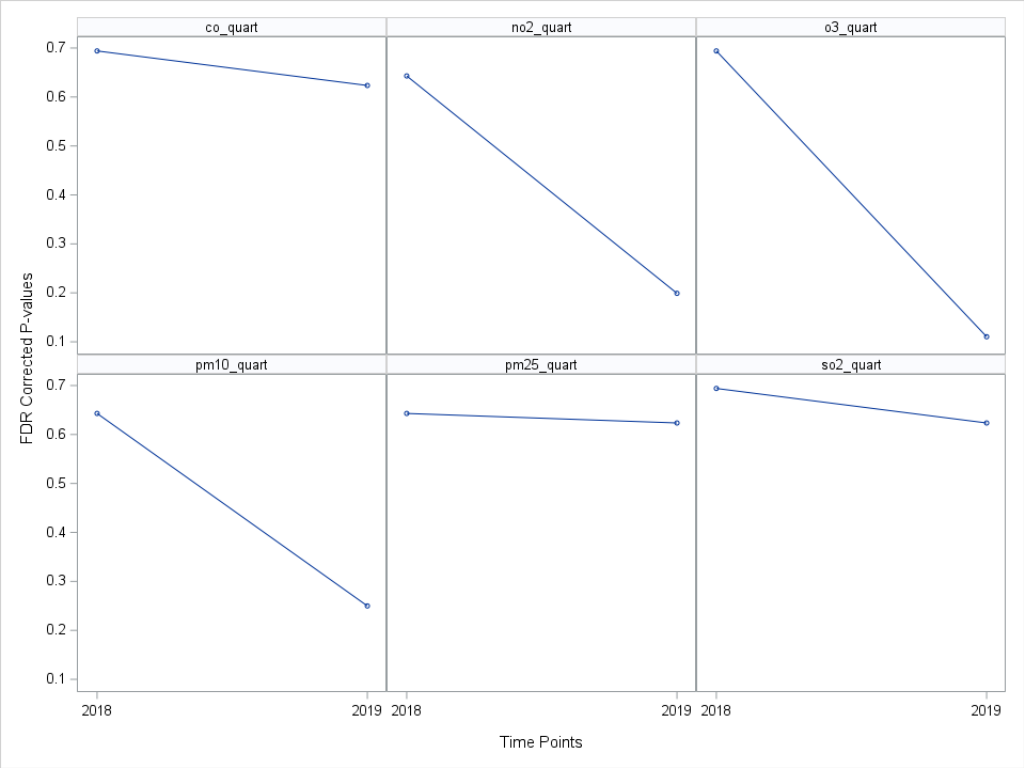
Conclusion

We conclude that there exists a significant spatial autocorrelation in prostate cancer deaths data and statistical models accounting for such correlations reveal that provinces with higher humidity report higher levels of prostate cancer deaths. More comprehensive prospective studies targeting more granular data, ideally at the individual level, are needed to describe the mechanism of action between environmental factors and prostate cancer diagnosis and deaths.

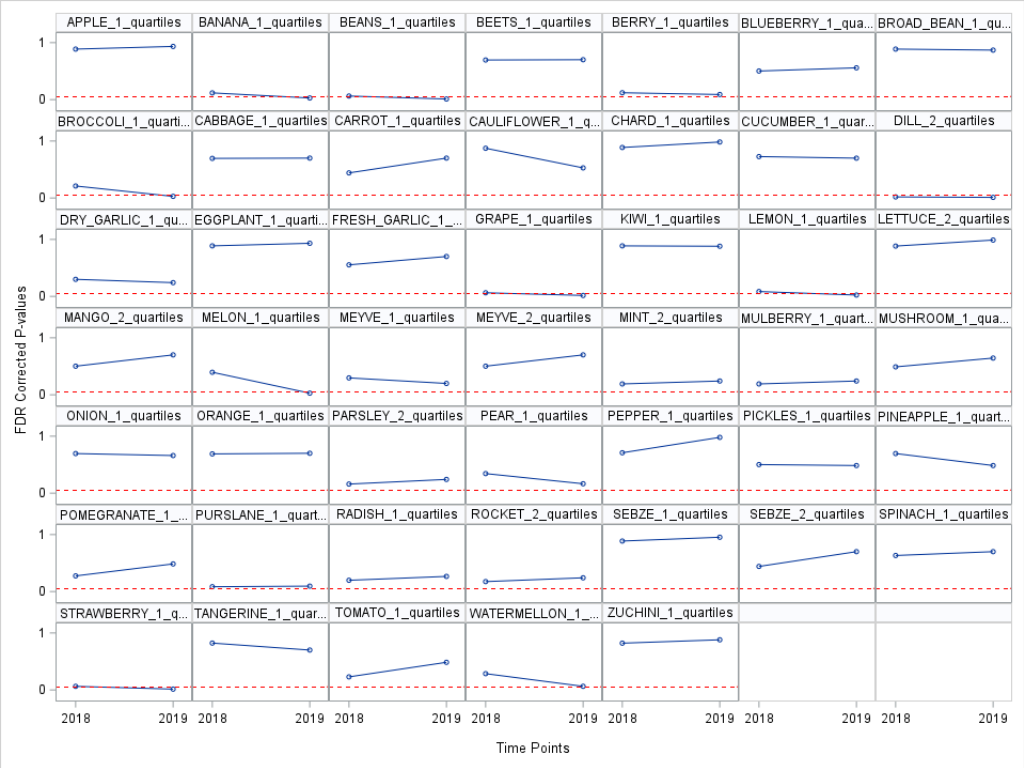
References

1. Global Burden of Disease Cancer Collaboration. 2016. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study. JAMA Oncol 3: 524–548.
2. Ferlay J, Ervik M, Lam F, et al. Global cancer observatory: cancer today. International Agency for Research on CancerLyon, France https://gco.iarc.fr/today 2018.
3. Sharma R. The burden of prostate cancer is associated with human development index: evidence from 87 countries, 1990–2016. EPMA Journal. 2019 Jun;10(2):137-52.
4. Luo LS, Jiang JF, Luan HH, Zi H, Zhu C, Li BH, Zeng XT. Spatial and temporal patterns of prostate cancer burden and their association with Socio-Demographic Index in Asia, 1990–2019. The Prostate. 2021 Oct 18.
5. Hsing AW, Sakoda LC, Chua S Jr. Obesity, metabolic syndrome, and prostate cancer. Am J Clin Nutr 2007; 86: s843–s857.
6. Platz EA, Giovannucci E. Prostate cancer. In: Cancer epidemiology and prevention. Oxford University Press; 2006. p. 1128–50.
7. Bashir MN. Epidemiology of prostate cancer. Asian Pac. J. Cancer Prev.2015; 16: 5137–41.
8. Zhu Y, Wang HK, Qu YY, Ye DW. Prostate cancer in East Asia: evolvingtrend over the last decade. Asian J. Androl. 2015; 17:48–57
9. Bailey, T. C., & Gatrell, A. C. (1995). Interactive spatial data analysis. Harlow, UK: Addison Wesley Longman Limited.
10. Tobler, W. R. (1970). A Computer Model Simulating Urban Growth in the Detroit Region. Economic Geography, 46: 234-240.
11 Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950 Jun 1;37(1/2):17-23.
12 Jones BL, Nagin DS. Proc TRAJ: A SAS Procedure for Group-Based Modeling of Longitudinal Data. InAnnual Meeting Home 2007.
13 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995 Jan;57(1):289-300.
14. JI GG, MP MM. Will the climate change affect the mortality from prostate cancer?. Archivos espanoles de urologia. 2007 Mar 1;60(2):119-23.
15. St-Hilaire S, Mannel S, Commendador A, Mandal R, Derryberry D. Correlations between meteorological parameters and prostate cancer. International journal of health geographics. 2010 Dec;9(1):1-1.
16. Eyles H, Jiang Y, Mhurchu CN. Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. Journal of the American Dietetic Association. 2010 Jan 1;110(1):106-10.

SUPPLEMENTARY MATERIALS



Supplementary Figure 1. Significance of Air-Quality Markers controlling for Average Humidity.



Supplementary Figure 2. Univariable Significance of Fruit and Vegetable Consumption Markers

KS4

Avrupa Birliği Üye Ülkeleri ve Aday Ülke Türkiye'nin Sağlık Göstergeleri Bakımından Çok Değişkenli İstatistiksel Yöntemlerle Analiz Edilmesi

Anıl TUZCU[1], Cengiz BAL[1]

[1]*Eskişehir Osmangazi Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Eskişehir*

Giriş: Sosyo-ekonomik göstergelerden biri olan sağlık hizmetleri, ülkelerin gelişmişlik düzeylerini belirlemek için kullanılan önemli bir kriterdir. Bu araştırmada Türkiye'nin sağlık değişkenlerinin, Avrupa Birliği'ne (AB) üye ülkeler ile karşılaştırılması ve konumunun belirlenmesi amaçlanmıştır.

Yöntem: Türkiye'nin AB'ye katılım sürecinde sağlık değişkenleri bakımından sıralama ve kümelemenin yapılmasının sağlık planlayıcılarına faydalı olacağı düşünülmüştür. Bu çalışmada, Türkiye ve AB'ye üye 27 ülkenin sağlık hizmetlerinin gelişmişlik düzeylerini belirmek amacıyla uluslararası kuruluş ve yayınların veri tabanlarından elde edilmiş güncel 17 sağlık değişkeni, çok değişkenli istatistiksel analizlerde kullanılmıştır. Değişkenlerin bağımlılık yapısını gidermek için temel bileşenler analizi, elde edilen faktör skorlarına göre karşılaştırmak için faktör analizi ve elde edilen skorlar ile ülkeleri kümelere ayırmak için kümeleme analizi kullanılmıştır. Belirlenen sağlık değişkenleri, Dünya Sağlık Örgütü (DSÖ) tarafından tanımlanmış küresel sağlık gözlemevi başlığı altında yer alan sağlık değişkenlerinden seçilmiştir. Araştırmada tercih edilen her bir sağlık değişkeni güncel verilerden (2017-2020 yılları) elde edilmiştir. Bu çalışmadaki, verilerin elde edilmesinde DSÖ, Avrupa İstatistik Ofisi (Eurostat), Ekonomik Kalkınma ve İş Birliği Örgütü (OECD), Dünya Bankası (The World Bank) ve T.C. Sağlık Bakanlığı Sağlık İstatistik Yıllığı veri tabanlarından yararlanılmıştır.

Bulgular: Ülkelerin gelişmişlik düzeylerini belirlemek için analizde kullanılacak 28 ülkeye ait 17 sağlık değişkeninden dört tane temel bileşenin öz değeri birden büyük olduğu anlaşıldığından dört faktöre indirgenmiştir. Bu bileşenler, toplam açıklanan varyans oranının %82,227'sini açıklamaktadır. Faktör grubun daha iyi yorumlanabilmesi için ortogonal döndürme yöntemlerinden varimax yöntemi kullanılmıştır.

Birinci faktör beş değişkenden (Bebek Ölüm Oranı, Beş Yaş Altı Ölüm Oranı, Yeni Doğan Ölüm Hızı, Ölü Doğum Oranı ve Ergen Doğurganlık Oranı), ikinci faktör altı değişkenden (Kaba Ölüm Oranı, Hastane Yatakları, Toplam Yaşam Beklentisi, Yıllık Nüfus Artış Oranı, Yetişkin Ölüm Oranı ve Doğuşta Beklenen Yaşam Süresi), üçüncü faktör üç değişkenden (Tüberküloz İnsidansı, Tüberküloza Bağlı Ölüm Oranı ve Anne Ölüm Oranı), dördüncü faktör üç değişkenden (Toplam Sağlık Harcamaları, Hemşire Personeli ve Kentsel Nüfus Oranı) oluşmuştur.

Kümeleme analizi ile homojen ülke sınıflandırılması amaçlanmıştır. Hiyerarşik olmayan kümeleme yöntemlerinden K-ortalamalar kümeleme yöntemi yardımıyla üç kümeye ayrılarak, üç grup oluşturulmuştur.

Değişken No	Sağlık Göstergeleri
D1	Bebek Ölüm Oranı
D2	Doğuşta Beklenen Yaşam Süresi
D3	Ergen Doğurganlık Oranı
D4	Hastane Yatakları
D5	Kaba Ölüm Oranı
D6	Kentsel Nüfus Oranı
D7	Toplam Yaşam Beklentisi
D8	Yeni Doğan Ölüm Hızı
D9	Yetişkin Ölüm Oranı (15-60 Yaş)
D10	Yıllık Nüfus Artış Oranı (%)
D11	Beş Yaş Altı Ölüm Oranı
D12	Tüberküloza Bağlı Ölüm Oranı (HIV-)

D13	Tüberküloz İnsidansı
D14	Ölü Doğum Oranı
D15	Hemşire Personeli
D16	Anne Ölüm Oranı
D17	Toplam Sağlık Harcamaları (GSYİH)

Bebek Sağlığı Faktörü
Bebek Ölüm Oranı (D1)
Beş Yaş Altı Ölüm Oranı (D11)
Yeni Doğan Ölüm Hızı (D8)
Ölü Doğum Oranı (D14)
Ergen Doğurganlık Oranı (D3)

Toplum Faktörü
Kaba Ölüm Oranı (D5)
Hastane Yatakları (D4)
Toplam Yaşam Beklentisi (D7)
Yıllık Nüfus Artış Oranı (D10)
Yetişkin Ölüm Oranı (D9)
Doğuşta Beklenen Yaşam Süresi (D2)

Hastalık Yükü ve Ölüm Faktörü
Tüberküloz İnsidansı (D13)
Türberküloza Bağlı Ölüm Oranı (D12)
Anne Ölüm Oranı (D16)

Sağlık Finansmanı ve Demografi Faktörü
Toplam Sağlık Harcamaları (D17)
Hemşire Personeli (D15)
Kentsel Nüfus Oranı (D6)

Sonuç: Çalışmada kullanılan sağlık değişkenlerine uygulanan analiz yöntemlerinden elde edilen sonuçlar; birinci, ikinci ve dördüncü faktörde Türkiye'nin üye ülkelere yakın olan değişkenleri olsa da henüz Avrupa Birliği üye ülkeleri seviyesine ulaşamadığını göstermiştir. Üçüncü faktörde elde edilen sonuçlara göre uygulanan çok değişkenli istatistiksel yöntemler ile Türkiye'nin birçok ülkeden daha iyi seviyede olduğu ve bu faktör ile AB seviyesini yakaladığı ortaya konmuştur. Üçüncü faktörde yer alan sağlık değişkenleri açısından, Türkiye'nin gelişmiş AB üye ülkeler seviyesinde olduğu görülmektedir. Birinci, ikinci ve dördüncü faktörde yer almakta olan sağlık değişkenleri bakımından, ülke olarak bu değişkenler üzerinde durulmalı ve değişkenlerin oranı gelişmiş ülkelerin seviyesine çekilmelidir.

KS5

Gözlemsel Çalışmalarda Yapılan Meta-Analiz Çalışmalarının Raporlanması

MOOSE (Meta-analysis of Observational Studies in Epidemiology) Kontrol Listesinin Türkçe'ye Uyarlanması

Arzu Baygül Eden[1], Neslihan Gökmen İnan[2]

[1]Koç Üniversitesi Tıp Fakültesi, Biyoistatistik Bölümü, İstanbul
[2]İstanbul Teknik Üniversitesi, Temel Bilimler, İstanbul

ÖZET

Amaç

Meta analizler kanıta dayalı tıp alanında gücü en yüksek çalışmalar olarak belirlenmiştir. Gözlemsel çalışmalar gerçek hayata en yakın veriyi sunan, bir tedaviyi ya da hastalığı kendi doğasında gözlemlemeyi sağlayan çalışmalardır. Gözlemsel çalışmaların meta analizinin raporlanmasında randomize kontrollü çalışmalarda olduğu gibi bir rehber ya da bir kontrol listesi üzerinden gidilmesi bilimsel ve standart bir yaklaşım olacaktır. Bu amaçla 1997 yılında Atlanta'da gözlemsel çalışmaları raporlama konusunun çalışıldığı bir çalıştay düzenlenmiş ve MOOSE (Meta-analysis of observational studies) (Meta-analysis of observational studies in epidemiology: The MOOSE (Meta-analysis of Observational Studies in Epidemiology) statement (JAMA 2000;283:2008–2012). raporlama kontrol listesi oluşturulmuştur. Bu çalışma MOOSE'da belirtilen kriterlerin dilimize kazandırılması ve araştırıcılar tarafından daha etkin bir şekilde kullanılarak gözlemsel çalışmaların meta analizinin daha sağlıklı şekilde yürütülmesini amaçlamaktadır.

Gereç ve Yöntemler

MOOSE kontrol listesi alanında uzman iki araştırıcı tarafından birbirine kör olarak Türkçe'ye tercüme edilmiştir. Araştırıcılar bir araya gelerek çevrilen dökümanları inceleyerek ortak bir dil üzerinde karar vermiştir. Uzman bir çevirmen tarafından yapılan ters çeviri işlemiyle Türkçe'ye çevrilmiş olan MOOSE ile tekrar İngilizce'ye çevrilmiş orijinal İngilizce versiyon arasındaki farklılıklar incelenmiştir. Kontrol listesinin Türkçe olarak kullanılabilirliğini araştırmak amacıyla seçilen ve Türkiye verileri üzerinde çalışılmış 10 gözlemsel meta analiz çalışması makalesi 2 araştırıcı tarafından MOOSE kontrol listesi kullanılarak, her bir madde açısından puanlanmıştır. Maddenin gerçekleşmesi durumunda "1" puan verilmiştir. Araştırıcılar arası uyum incelenmiştir.

Ayrıca makalelerin toplam puanları hesaplanmış, her iki değerlendiricinin her bir maddeye verdiği puanlar arası uyum Cronbach Alpha katsayısıyla incelenmiştir.

Bulgular

Her bir maddede değerlendiriciler arası uyum oldukça iyi düzeyde çıkmıştır (Uyum oranı 0,80-1,00).

Toplam puanlar arasında yüksek düzeyde uyum olduğu görülmüştür. (Cronbach Alpha Katsayısı 0,939)

Sonuç

MOOSE kontrol listesinin Türkçe versiyonu Türk araştırmacılar tarafından gözlemsel çalışmalar için uygulanan meta analizlerde kullanılabilecek uygun bir araçtır.

Anahtar Sözcükler: MOOSE, Gözlemsel çalışma, Meta analiz, Kontrol listesi

Referanslar

1. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000 Apr 19;283(15):2008-12. doi: 10.1001/jama.283.15.2008. PMID: 10789670
2. Mehmet Birhan Yılmaz, Mustafa Kılıçkap, Adnan Abacı, Cem Barçın, Fahri Bayram, Doruk Karaaslan, Hüseyin Göksülük, Meral Kayıkçıoğlu, Necla Özer, Gültekin Süleymanlar, Mahmut Şahin, Lale Tokgözoğlu,

İlhan Satman. Temporal changes in the epidemiology of diabetes mellitus in Turkey: A systematic review and meta-analysis. Turk Kardiyol Dern Ars. 2018; 46(7): 546-555

3. Dilek Ural, Mustafa Kılıçkap, Hüseyin Göksülük, Doruk Karaaslan, Meral Kayıkçıoğlu, Necla Özer, Cem Barçın, Mehmet Birhan Yılmaz, Adnan Abacı, Şule Şengül, Turgay Arınsoy, Yunus Erdem, Yavuz Sanisoğlu, Mahmut Şahin, Lale Tokgözoğlu. Data on prevalence of obesity and waist circumference in Turkey: Systematic review, meta-analysis and meta-regression of epidemiological studies on cardiovascular risk factors. Turk Kardiyol Dern Ars. 2018; 46(7): 577-590

4. Görkem Karakaş Uğurlu, Mustafa Uğurlu. The Prevalence of Sexual Dysfunctions in Women with Diabetes and Its Relationship with Diabetic and Demographic Factors: A Meta-analysis and Meta-regression Study. Ankara Med J. 2020; 20(4): 798-813

5. Medine Alpdemir, Mehmet Fatih Alpdemir. Vitamin D deficiency status in Turkey: A meta-analysis, Int J Med Biochem 2019;2(3):118-31

6. Çetinkaya Ü, Caner A. Prevalence of Microsporidiosis in Different Hosts in Turkey: A Meta-analysis. Turkiye Parazitol Derg 2020;44(4):232-8.

7. Mustafa Güzel, Orhan Akpınar, Muhammet Burak Kılıç. Prevalence of Rotavirus-Associated Acute Gastroenteritis Cases in Early Childhood in Turkey: Meta-Analysis. Children (Basel). 2020 Oct; 7(10): 159.

8. Neslihan Keser Özcan , Nur Elçin Boyacıoğlu, Hüsniye Dinç. Postpartum Depression Prevalence and Risk Factors in Turkey: .A Systematic Review and Meta-Analysis. Archives of Psychiatric Nursing 31 (2017) 420–428

9. Necla Özer, Mustafa Kılıçkap, Lale Tokgözoğlu, Hüseyin Göksülük, Doruk Karaaslan, Meral Kayıkçıoğlu, Mehmet Birhan Yılmaz, Cem Barçın, Adnan Abacı, Mahmut Şahin. Data on smoking in Turkey: Systematic review, meta-analysis and meta-regression of epidemiological studies on cardiovascular risk factors. Turk Kardiyol Dern Ars. 2018; 46(7): 602-612

10. Zekiye Karaçam & Demet Çelİk (2021) The prevalence and risk factors of gestational diabetes mellitus in Turkey: a systematic review and meta-analysis, The Journal of Maternal-Fetal & Neonatal Medicine, 34:8, 1331-1341, DOI: 10.1080/14767058.2019.1635109

11. Meral Kayıkçıoğlu, Lale Tokgozoglu, Mustafa Kılıçkap, Hüseyin Göksülük, Doruk Karaaslan, Necla Özer, Adnan Abacı, Mehmet Birhan Yılmaz, Cem Barçın, Kenan Ateş, Fahri Bayram, Mahmut Şahin, Dilek Ural. Data on prevalence of dyslipidemia and lipid values in Turkey: Systematic review and meta-analysis of epidemiological studies on cardiovascular risk factors. Turk Kardiyol Dern Ars. 2018; 46(7): 556-574

KS6

Örneklemin Evrendeki Oranı ile Güven Aralığı Genişliğinin Uyumsuzluğu

Mehmet Güven GÜNVER[1], **Aydan MALÇOK[4],** Uğurcan SAYILI[2,5], Mine SEZGİN[3,5], Hatice Sena ÖNER[4] , Sevda ÖZEL YILDIZ[1], Başak GÜRTEKİN[1], Eray YURTSEVEN[1]

[1] İstanbul Üniversitesi, İstanbul Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, İstanbul, Türkiye
[2] Karaköprü İlçe Sağlık Müdürlüğü, Şanlıurfa, Türkiye
[3] İstanbul Üniversitesi, İstanbul Tıp Fakültesi, Nöroloji Anabilim Dalı, İstanbul, Türkiye
[4] İstanbul Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik ve Tıp Bilişimi Doktora Programı, İstanbul, Türkiye

Giriş: Bu çalışmanın amacı, örneklemin evreni temsil oranı ile güven aralığının genişliği arasındaki ilişkiyi değerlendirmektir.

Yöntem: Örneklem büyüklüklerini belirlemek amacıyla sürekli veriler için $n = \frac{Nt^2x^2}{S^2(N-1)+t^2x^2}$ ve oransal veri için $n = \frac{Nt^2(pq)}{S2(N-1)+t^2(pq)}$ formülü kullanıldı. N:popülasyon büyüklüğü; n:örneklem büyüklüğü; x: ortalama; S:Standart sapma, p: incelenen olayın görülme sıklığı; q:incelenen olayın görülmeme sıklığı; t: Belirli serbestlik derecesinde ve saptanan yanılma düzeyinde t tablo-sundan bulunan teorik değeri ifade etmektedir. $st(e) = \frac{S}{\sqrt{n}}$ ve $st(e) = \frac{(p*q)}{\sqrt{n}}$ formülü ile hesaplandı. St(e), standart hatayı ifade etmektedir. Güven aralıkları $x \pm z_{a2}$*st(e) olarak belirlendi. Örneklem büyüklüğünü belirlemek amacıyla literatürden örnek çalışmalar kullanıldı. Örneklem ve güven aralığı hesabında kullanılan veriler; sürekli değişken için: x±S=138,4±19,45; x±S=138,4±19,45; oransal veri için p:%18, d:0,02. İstatistiksel analizler Microsoft Office Excel 2016 ile yapıldı.

Bulgular: Ortalaması 138,4 standart sapması 19,45 olan sürekli veri için değerlendirildiğinde n/N oranı 16.3% iken CI: 135,41-141,39; n/N oranı 6,1% iken CI: 135,59-141,22 olarak giderek daralmakta ve n/N: 0,195% iken CI: 135,67-141,13 olmaktadır. Örneklem büyüklüğü 100.000'in üzerine çıktığında güven aralığı artmadığı; örneklem büyüklüğü 8.500.000.000 olduğunda bile n/N oranı 0,0000023% iken CI: 135,67-141,13 olarak en dar halinde kalmaktadır. Ortalamayı sabit tutup standart sapmayı 2 olarak belirlediğimizde n/N oranı %94,9 iken CI: 138,27-138,53; n/N oranı 86% iken CI: 138,32-138,48 olarak giderek daralmaktadır. Ortalamayı sabit tutup standart sapmayı 50 olarak aldığımızda ise n/N oranı 2,9% olduğunda da 0,00003% olduğunda da CI: 120,20-156,60 olarak hesaplanmış ve tüm değerler için aynı güven aralığı elde edilmiştir.

P=0,18 d=0,02 olan oransal veri için değerlendirildiğinde evren büyüklüğü 1000 iken n/N:59,20% olup CI: %15,37-%21,63; evren büyüklüğü 10.000 iken n/N: 12,65%'e azalmasına rağmen CI: %16,36-%20,64 daraldığı görüldü. Evren büyüklüğü 1.000.000 iken n/N:0,1446% olup CI: %16,50-%20,50; evren büyüklüğü 1.000.000'un üzerine çıktığında ise n/N düşmesine rağmen CI sabit kaldığı görüldü. P=0,40 d=0,02 olan oransal veri için n/N oranı 69,8% CI: %38,22-%41,78; n/N oranı 0,23% CI: %39,02-%40,98 olarak hesaplanmış ve n/N oranı azalmasına rağmen CI sabit kalmıştır. P=0,4 d=0,2 olan veri için n/N oranı 2,6% CI: %35,29-%44,71; n/N oranı 0,90% CI: %35,68-%44,62 olarak hesaplanmıştır. n/N oranı azalmasına rağmen CI sabit kalmıştır.

Sonuç: Evren büyüklüğünün aşırı artmasına rağmen örneklem büyüklüğü görece daha az bir artış göstermekte ve evren büyüklüğü artarken n/N oranı giderek azalmaktadır. Güven aralığı formülü hesabında evren büyüklüğü olmamasına bağlı olarak örneklem büyüklüğü aşırı artarken n/N azalmasına rağmen güven aralığı genişliği daralmaktadır. Güven aralığı formülüne evren veya örneklemin evren temsiliyeti boyutu olmaması yanıltıcı sonuçlara yol açabilir. Bu konuda yapılacak ileri çalışmalara ihtiyaç vardır.

Anahtar Sözcükler: Güven Aralığı, Örneklem Büyüklüğü, Popülasyon

N	n	n/N	result	%95 CI
1000	163	16,30%	138,4	135,41-141,39
3000	183	6,10%	138,4	135,59-141,22
5000	188	3,76%	138,4	135,62-141,18
10000	191	1,91%	138,4	135,64-141,16
100000	195	0,20%	138,4	135,67-141,13
1000000	195	0,02%	138,4	135,67-141,13
10000000	195	0,00%	138,4	135,67-141,13
8500000000	195	0,00%	138,4	135,67-141,13

Not: **Ortalama ± st sapma: 138,4±19,45**

N	n	n/N	result	%95 CI
1000	949	94,90%	138,4	138,27-138,53
3000	2579	86,00%	138,4	138,32-138,48
5000	3931	78,60%	138,4	138,34-138,46
10000	6478	64,78%	138,4	138,35-138,45
100000	15538	15,54%	138,4	138,37-138,43
1000000	18063	1,81%	138,4	138,37-138,43
10000000	18363	0,18%	138,4	138,37-138,43
8500000000	18396	0,00%	138,4	138,37-138,43

Not: **Ortalama ± st sapma: 138,4±2**

N	n	n/N	result	%95 CI
1000	29	2,90%	138,4	120,20-156,60
3000	29	0,97%	138,4	120,20-156,60
5000	29	0,58%	138,4	120,20-156,60
10000	29	0,29%	138,4	120,20-156,60
100000	29	0,03%	138,4	120,20-156,60
1000000	29	0,003%	138,4	120,20-156,60
10000000	29	0,0003%	138,4	120,20-156,60
8500000000	29	0,00%	138,4	120,20-156,60

Not: **Ortalama ± st sapma: 138,4±50**

N	n	n/N	result	%95 CI
1000	592	59,20%	18,50%	%15,37-%21,63
3000	977	32,57%	18,50%	%16,07-%20,93
5000	1123	22,46%	18,50%	%16,23-%20,77

10000	1265	12,65%	18,50%	%16,36-%20,64
100000	1428	1,43%	18,50%	%16,49-%20,51
1000000	1446	0,14%	18,50%	%16,50-%20,50
10000000	1448	0,01%	18,50%	%16,50-%20,50
8500000000	1448	0,00%	18,50%	%16,50-%20,50

Not: **p:%18, d:0,02**

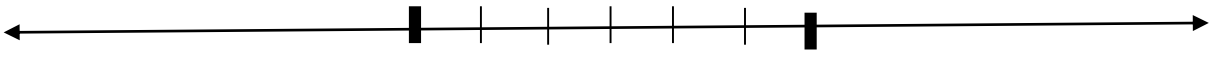
N	**n**	**n/N**	**result**	**%95 CI**
1000	698	69,80%	40,00%	%38,22-%41,78
3000	1304	43,46%	40,00%	%38,70-%41,30
5000	1578	31,56%	40,00%	%38,82-%41,18
10000	1874	18,74%	40,00%	%38,91-%41,09
100000	2254	2,25%	40,00%	%39,01-%40,99
1000000	2300	0,23%	40,00%	%39,02-%40,98
10000000	2305	0,02%	40,00%	%39,02%40,98
8500000000	2305	0,00%	40,00%	%39,02-%40,98

Not: **p:%40, d:0,02**

N	**n**	**n/N**	**result**	**%95 CI**
1000	26	2,60%	40,00%	%35,29-%44,71
3000	27	0,90%	40,00%	%35,38-%44,62
5000	27	0,54%	40,00%	%35,38-%44,62
10000	27	0,27%	40,00%	%35,38-%44,62
100000	27	0,03%	40,00%	%35,38-%44,62
1000000	27	0,003%	40,00%	%35,38-%44,62
10000000	27	0,0003%	40,00%	%35,38-%44,62
8500000000	27	0,0000003%	40,00%	%35,38-%44,62

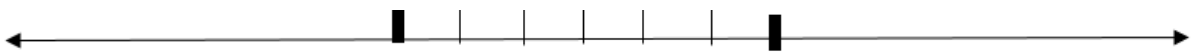
Not: **p:%40, d:0,2**

(n/N=16,3%)



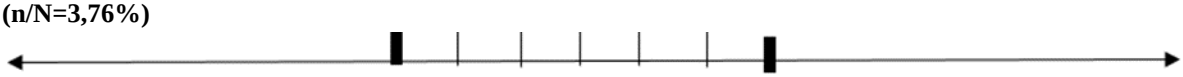
(135,41 141,39)

(n/N=6,3%)



(135,59 141,22)

(n/N=3,76%)



(135,62 141,18)

Kaynaklar

1. Altman DG: Confidence intervals in practice. In: Altman DG, Machin D, Bryant TN, Gardner MJ. BMJ Books 2002; 6–9.
2. Bulpitt CJ: Confidence intervals. Lancet. 1987;1:494–497
3. Gardner MJ, Altman DG: Confidence intervals rather than P-values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746–50.

KS7

Comparison of Quantile Regression And Linear Regression Method: A Simulation Study

Didem Ünal[1], Gülhan Temel[2]

[1]Turkish Red Crescent, Ankara,
[2]Mersin University Department of Biostatistics and Medical Informatics, Mersin

Introduction: Regression analysis, which is frequently used in health sciences, examines the relationship between independent variable(s) and dependent variable. The Quantile Regression Method is a powerful regression analysis method. The aim of this study is to compare the Quantile Regression Method in cases where there are outliers in different sample sizes.

Method: A regression equation is established with the help of the relations between the variables and model estimation is made with this established regression equation. There are multiple regression methods that can be used for model estimation. In order to make the results reliable, it is necessary to choose the method that best fits the model among these regression methods. The most commonly used regression method is the Least Squares Method. In order to use OLS (Least Squares) method, some assumptions are needed in the data set. There are some situations in which these assumptions cannot be met while performing the regression analysis. The most common distortion of assumption is assumption of normality. When the assumption of normality could not be achieved, it also often leads to the problem of varying variance. In such cases, data transformation techniques are generally used. The Quantile Regression Method, which was developed as an alternative to the OLS method, which can be done without using data transformation techniques, is a powerful regression analysis method, especially in cases where distortion of assumptions that are frequently encountered in health data. The Quantile Regression Method does not require any assumptions on the data and provides a strong estimation for both parametric and non-parametric distributions compared to the OLS method. In other words, although the OLS method is sensitive to some deviations in the data set, the QR method is not. In this study, the performances of OLS and QR (Quantile Regression) methods were evaluated by generating outliers in different sample sizes. Simulation study;

- Sample size n=30, 100 and 1000
- %5, %10 and %20 Outlier
- For OLS Method and QR tau=0.25, 0.50, 075

was made to be. First, the results were evaluated by making analyzes in the absence of outliers. Then, in the presence of outliers, the results were re-evaluated, the performances of the OLS and QR methods were examined.

Results: In line with the criteria specified for the simulation study, quantile regression analyzes were performed in the R Studio Package Program. In all analyzed sample sizes, the QR results were examined, and it was seen that the R^2 explanation percentage in the model had the highest ratio for tau=0.75 in the datasets with 5% and 10% outliers. In the datasets with 20% outliers, R^2 explanation percentage has the highest ratio for tau=0.25. In the case where outlier observation is not assigned, it is seen that the R^2 explanation percentage has the highest ratio for tau=0.75 when the sample size is 30 and 100.

Discussion and Conclusion: Instead of the OLS method, which deals with only the average of the data, it is more reliable to use QR method, which also deals with the results based on the median and some quarterly measures of the data in cases where the data obtained in many studies in the field of health do not fit with the assumption of normality.

Keywords: Linear Regression Analysis, Quantile Regression, Minimum Absolute Deviation, Varying Variance, Normality Assumptionf

References

1. Koenker, R., “Quantile Regression, Econometric Society Monographs”, Cambridge University Press, 2005.
2. Hao L., Naiman D.Q., “Quantile Regression”, Sage Publications, Inc, California, 2013

3. Sümbüloğlu V., Alpar R. , Özdemir P.; "Değişkenler arası ilişkilerin incelenmesi", Hacettepe Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara
4. Koenker R., "This article has been prepared for the statistics section of the International Encyclopedia of the Social Sciences edited by Stephen Fienberg and Jay Kadane", The research was partially supported by NSF Grant, Version: October 25, 2000.
5. Koenker R., Machado J.A.F., "Goodness of Fit and Related Inference Processes for Quantile Regression", Journal of the American Statistical Association , 1999, ISSN: 0162- 1459
6. Petscher Y., Logan J.A.R., "Quantile Regression in the Study of Developmental Sciences", Development Methods, 2014 May-Jun;85(3):861-881
7. Chen C.L., "An Introduction to Quantile Regression and the QUANTREG Procedure", SAS eranceInstitute Inc., Cary, NC, 2005, Paper 213-30
8. Baum C.F., "Quantile regression", Applied Econometrics Boston College, Spring 2013

KS8

Determination of Effective Degrees of Freedom of Parameters in Box-Cox Cole Green (BCCG) Distribution and Comparison of Model Performances Smoothed for Pediatric Growth Curves

Eda Çakmak[1], Ergun Karaağaoğlu[2]

[1]*Başkent University Faculty of Health Science, Ankara*
[2]*Hacettepe University Faculty of Medicine, Department of Biostatistics, Ankara*

Objective: Growth curves are one of the important indicators used for observing child's development, health and nutritional status. In the growth stages of children anthropometric measurements generally show a skewed distribution because the increase with age is not constant, and one of the most frequently used methods to remove skewness is the LMS(Lambda, Mu, Sigma) method. It is very fundamental to determine the effective degrees of freedom in order to obtain smooth curves in the LMS method. To that end, it is aimed to compare the model performances calculated from smoothed curves at different effective degrees of freedom with the LMS method.

Methods: The LMS method is based on the BCCG(ν, μ, σ) (Box-Cox Cole and Green) distribution. LMS method is defined by three parameters, the L curve(ν), the Box-Cox power transformation to convert the distribution to the normal distribution, the M curve(μ), the median value and the S curve (σ), the coefficient of variation. In this study, height and head circumference measurements in the "Dutch Boys" data set with "*dbhh*" package in *R* were used. Data comprises the height measurements of 2601 boys with 0-5 age range and the head circumference measurements of 2228 boys with 0-3 age range. Growth curves were constructed in RefCurv 0.4.2 program. LMS values of height and head circumference measurements of children in a specific age range were calculated. According to the BCCG distribution, model performances were compared with different effective degrees of freedom in the penalized spline(P-spline) and cubic spline methods after smoothing the data of height and head circumference. Global Deviance, AIC(Akaike's Information Criterion) and SBC(Schwarz Bayesian Criterion) values were considered in the evaluation of model performances.

Results: Whereby smoothed height and head circumference data in the LMS method which models only the skewness of the distribution, the effective degrees of freedom for the LMS curve were obtained as BCCG(2,8,4) and BCCG(2,7,4) respectively. As regards these values, Global Deviance=12846.25 AIC=12886.25 SBC=13003.53; Global Deviance=12849.88 AIC=12889.88 SBC=13007.16 for penalized spline and cubic spline in height data respectively, Global Deviance=8067.52 AIC=8105.52 SBC=8213.89; Global Deviance=8068.96 AIC=8106.96 SBC=8215.43 for penalized spline and cubic spline in head circumference data respectively were calculated.

Conclusion: Model performances of height and head circumference growth curves were found to be quite close to each other in the effective degrees of freedom calculated as a result of the penalized spline and cubic spline smoothing method in the BCCG distribution. It is very important to choose the effective degree of freedom that provides the best smoothing depending on the type and parameter of the distribution in the construction of the growth curves.

Keywords: Growth curves, BCCG distribution, LMS method, P-spline, Cubic spline

KS9

İstatistik ve Olasılık Dergilerinde Açık Erişim ve Abonelik Erişimli Dergiler: 1999-2018 dönemi için karşılaştırmalı bir inceleme

İlker ETİKAN

Yakın Doğu Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, Lefkoşa/KKTC

Özet

SCImago dergi indeksi, bir derginin kalitesinin ölçüsüdür. İndeks, hangi dergilerin en çok atıf aldığını, hangi dergilerin kalite ve bilimsel güvenilirliğin bir göstergesi olarak kullanıldığını belirtir. Bunu, her dergiye derginin SCI Journal Rank (SCI Impact Factor) olarak bilinen bir numara vererek yapar; sayı ne kadar yüksek olursa, derginin kalite göstergesi de o kadar yüksek olur. Bu çalışma, Biyoistatistik dergilerini de içeren İstatistik ve Olasılık alanındaki 35 Açık Erişim dergisi ve 166 Abonelik Tabanlı Dergiyi incelemektedir. Amaç, istatistik ve olasılık yayıncılığındaki ortak eğilimleri göstermek için hem açık erişim (OA= Open Access) hem de abonelik erişimi (SA= Subscription Access) dergileri için SCI etki faktörlerini karşılaştırmaktır. Bu çalışmanın verileri SCImago Journal Index and Country Ranking (SJR) web sitelerinden alınmıştır. Açık erişim ve Abonelik erişimi dergilerinin oranına bakmak ve her birinin Etki faktörlerindeki (IF) değişiklikleri değerlendirmek için üç istatistiksel yöntem kullanıldı. Bunlar; bağımsız gruplarda t-testi, ki-kare testi ve iki karşılık gelen regresyon modeli (two corresponding regression models) kullanılmıştır. Bu analizlerin bulguları, 2016-2018 döneminde açık erişim Olasılık ve İstatistik-Biyoistatistik dergilerinin Etki Faktörlerinin abonelik erişimli dergilere göre önemli ölçüde arttığını göstermektedir.

Anahtar Sözcükler: • Açık Erişim • Abonelik Erişimi • Biyoistatistik • SCImago dergi sıralaması göstergesi • Etki faktörü

Giriş

Araştırmacılar, kendi alanlarındaki bilgilere ulaşmak için çoğunlukla akademik dergileri kullanırlar. Akademisyenlerin yeni fikirler ve teoriler hakkında sohbet etmelerinin bir yolu olarak da kabul edilebilir. Gerçekten de, birçok alanda, iki veya daha fazla teorisyen arasında bir literatür akışı bulunabilir. Akademik dergiler, Açık erişim (OA= Open Access) veya Abonelik erişimli (SA= Subscription Access) olabilir. Açık erişim dergileri ücretsizdir ve bir abonelik veya ödeme olmaksızın indirilebilir ve bu nedenle kolayca erişilebilir durumdadır. Abonelik erişimli dergi makaleleri ya bireysel olarak ödenebilir ya da bir abonelik paketinin bir parçasını oluşturabilir (derginin tamamı veya dergi koleksiyonu için). Birçok araştırmacı, bilginin herkes için hazır ve kullanılabilir olması gerektiğine inandığından abonelik dergileri hakkında devam eden bir tartışma vardır (Van Noorden, 2013). Ayrıca, dergilerin açık erişim yayıncılığında ısrar etmesi yaygın bir politika haline gelmektedir (Van Noorden, 2013). Fikirlerin yayılması, alanlarda araştırmayı ilerletmenin anahtarıdır ve bu nedenle abonelik dergileri tartışmalı olarak birçok fikrin geniş çapta yayılmasını engeller.

Bir derginin geçerliliğini ve kalitesini değerlendirmek için kullanılan en yaygın parametre derginin Etki faktörüdür (IF= Impact Factor). Bir derginin Etki faktörü değeri, 1955 yılında Eugene Garfield tarafından geliştirilmiştir. Dergide yayınlanan makalelerin atıf sayısından hesaplanır (Garfield, 1972; Garfield, 2006). Bunu takiben, bir grup İspanyol araştırmacı, bir derginin kalitesini değerlendirmek için SCImago Journal Ranking (Garfield, 1972) olarak bilinen alternatif bir yöntem geliştirmişlerdir. Araştırmacılar giderek açık ve abonelik veri tabanlarında yayınlanan çalışmalara ve gelecekte her birinin olumlu ya da olumsuz yönlerini değerlendirmeye başlamışlardır (Kazıktaş ve diğerleri 2019 ve Polat ve diğerleri 2019). İstatistik ve olasılık alanları sınırlıdır ve bu nedenle bu çalışmanın odak noktası 1999-2018 dönemindeki etki faktörlerini karşılaştırmak ve analiz etmektir. Bu, özellikle açık erişim ve aboneliğe dayalı dergilerin kalitesi hakkında tartışmalara katılmış olabilecek araştırmacılar için ilginç olabileceğini düşünüyorum.

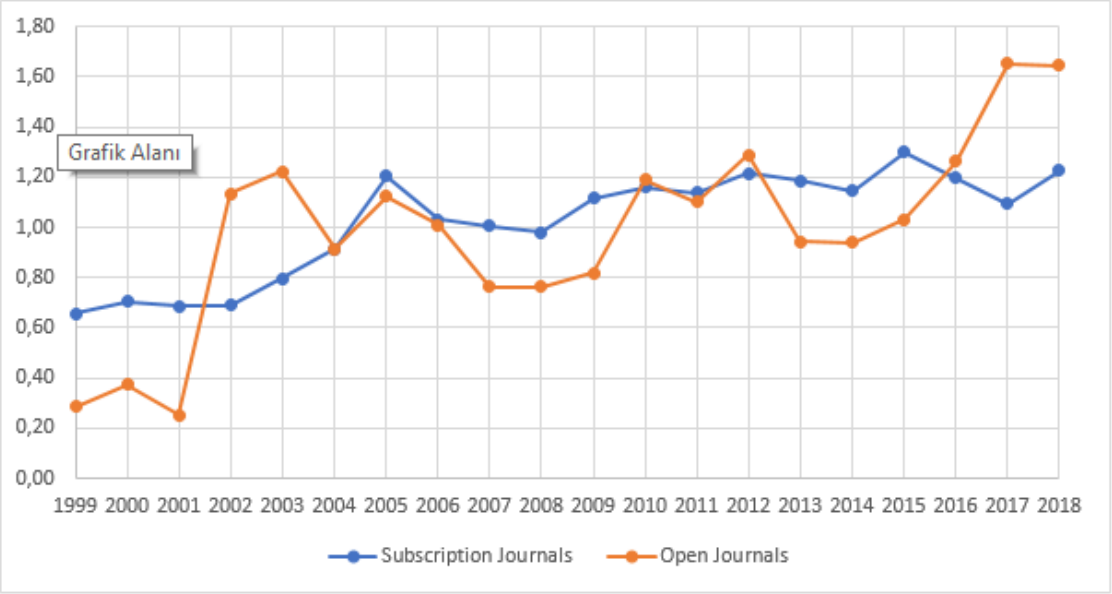
Gereç ve yöntemler

Bu çalışma için veriler SCImago Journal and Country Rank Database'den (SCImago web sitesinden) toplanmıştır. Analiz edilen verilerin tamamı SCImago veri tabanından alınmıştır. Açık Erişim (OA= Open Access) ve Abonelik Erişimi (SA= Subscription Access) dergilerinin sayısını değerlendirmek ve etki faktörlerindeki eğilimleri incelemek için üç istatistiksel analiz yapılmıştır. İlk olarak, 1999-2018 yılları arasında abonelik ve açık erişim dergilerinin ortalama Etki faktörleri (IF) değerlerini incelemek için bağımsız gruplarda t-testi uygulanmıştır.

Ki-kare bağımsızlık testi ise, aynı zamanda Abonelik erişimi ve Açık erişim dergilerinin sayı oranlarının belirtilen yıllar içinde önemli ölçüde değişip değişmediğini araştırmak için kullanılmıştır. Bunu takiben, Abonelik Erişimi ve Açık Erişim dergilerinin yıllar içindeki eğilimlerini (orantılı ve etki faktörü rakamları olarak) incelemek için Abonelik Erişimi ve Açık Erişim dergileri için ayrı ayrı iki karşılık gelen regresyon modeli (two corresponding regression models) uygulandı. Bağımlı değişkenler Abonelik Erişimi ve Açık Erişimler etki faktörleri ve yıllar bağımsız değişkenlerdir. Her dergi için ayrı bir değişken olarak yıllar seçilmiştir. Veriler istatistik yazılımı STATA 14 kullanılarak analiz edilmiştir.

Sonuçlar

Aşağıdaki Şekil 1'de, 1999 ve 2018 yılları arasındaki ortalama yıllık Etki faktörlerinin abonelik (SA= Subscription Access) ve açık erişim (OA= Open Access) dergilerine göre dağılımını göstermektedir. Genel olarak, hem açık erişimli hem de abonelik dergilerinin etki faktörlerinin ortalamaları 1999-2018 döneminde artmıştır. 1999 yılında abonelik dergilerinin ortalama etki faktörü 0,66 iken, açık erişim dergiler için ortalama 0,22 idi. 2002 ve 2004 yılları arasında, Açık Erişim dergilerinin Abonelik Erişimi dergilerine göre Etki faktörleri açısından yükselmeye başlamıştır. İlginç bir şekilde, 2018'de açık dergilerin ortalama etki faktörü (ortalama=1.64), abonelik dergilerinin ortalama etki faktöründen (ortalama=1.23) daha yüksek olduğu görünmektedir.



Şekil 1: Abonelik ve Açık erişim dergilerine göre ortalama yıllık etki faktörlerinin yıllar içindeki dağılımı

Tablo 1, Abonelik ve Açık erişim dergilerine göre 1999, 2005, 2010, 2015 ve 2018 yıllarında ortalama, Standart sapma ve Medyan etki faktörlerinin dağılımını göstermektedir.

Tablo 1. Abonelik ve Açık erişim dergilerine göre **1999 ile 2018** arası dağılımları

	Açık Erişimli Dergiler		Abonelik Erişimli Dergiler		
	Ortalama±S	Medyan	Ortalama±S	Medyan	pɪ
1999	0.28±0.19	0.20	0.66±0.58	0.52	0.21
2005	1.12±2.29	0.32	1.21±1.91	0.86	0.89
2010	1.19±1.38	0.58	1.16±0.98	0.90	0.90
2015	1.03±1.22	0.61	1.30±1.72	0.87	0.39
2018	1.64±2.12	0.96	1.22±0.96	0.95	0.07

ɪ: Bağımsız gruplarda t-testi

Abonelik erişimli (SA= Subscription Access) ve Açık erişimli (OA= Open Access) dergileri arasında beş farklı yılda istatistiksel fark olup olmadığını araştırmak için bağımsız gruplarda t-testi yapılmıştır. Tablo incelendiğinde, beş yıl boyunca Abonelik erişimli ve Açık erişimli dergilerin etki faktörleri arasında istatistiksel olarak anlamlı bir fark olmadığını görülmektedir ($p>0.05$).

Tablo 2: Açık Erişim ve Abonelik erişim dergilerinin toplam dergi sayısı içindeki oranları

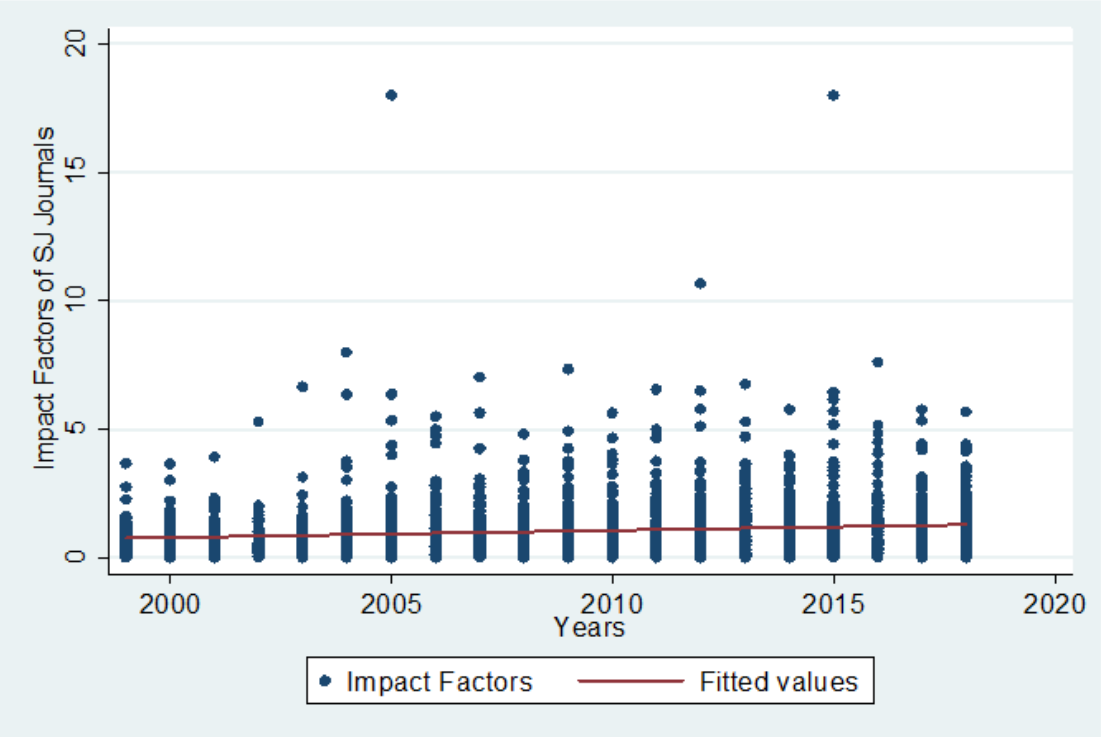
	Açık Erişimli Dergiler		Abonelik Erişimli Dergiler		Toplam		p
Yıllar	N	%	N	%	N	%	
1999	4	4.0	97	96.0	101	100	**0.001**
2018	34	17.1	165	82.9	199	100	
2005	10	8.5	111	91.4	121	100	**0.026**
2018	34	17.1	165	82.9	199	100	
2010	23	14.7	134	85.3	157	100	0.534
2018	34	17.1	165	82.9	199	100	
2015	33	17.4	157	82.6	190	100	0.941
2018	34	17.1	165	82.9	199	100	

Tablo 2, Olasılık ve İstatistik-Biyoistatistik dergileri için Abonelik Erişim ve Açık Erişim dergilerinin toplam sayısını göstermektedir. 1999'da 101 dergiden sadece 4'ü Açık Erişim dergi iken 97'si Abonelik Erişim dergisiydi. Açık Erişim dergilerinin artan sayılarını göstermek için bağımsız gruplarda ki-kare testi kullanıldı. Yukarıdaki tablodan (tablo 2) aşağıdaki hipotez oluşturulabilir:

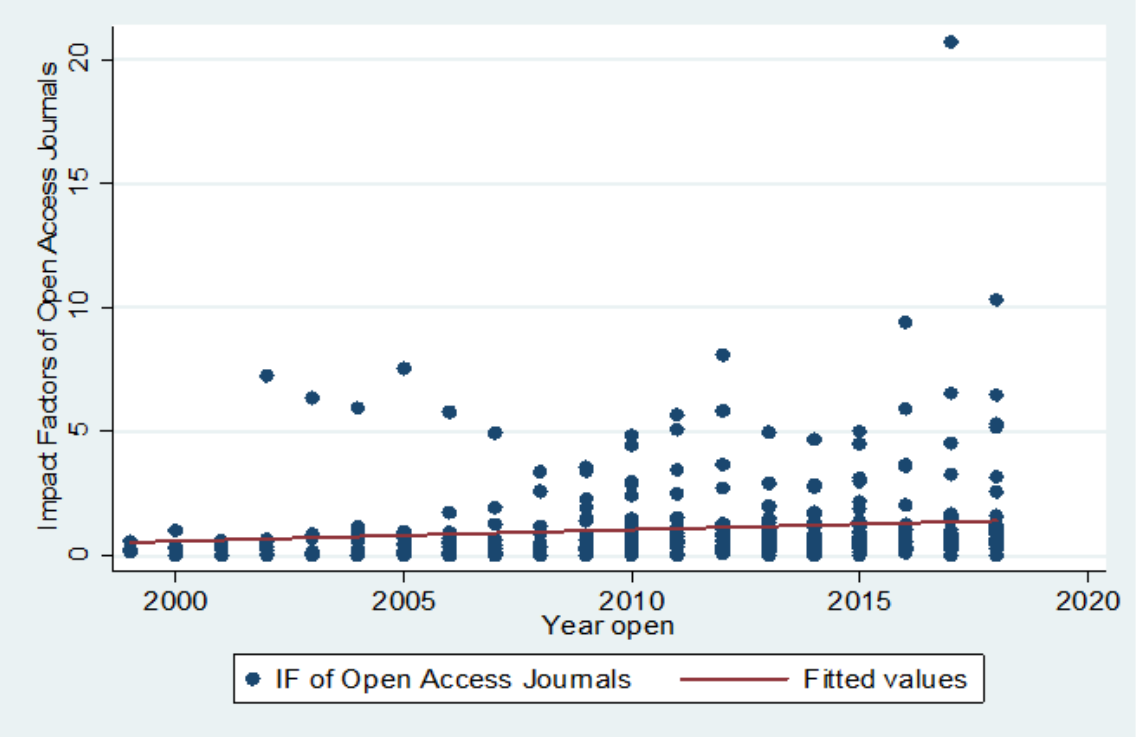
• Ho = p1 = p2 (1999-2018 döneminde Abonelik Erişim ve Açık Erişim dergilerinin sayı oranında fark yoktur);

• H1 = p1 ≠ p2 (1999-2018 döneminde Abonelik Erişim ve Açık Erişim dergilerinin sayı oranında bir fark vardır)

1999 ve 2018 yılları arasında Abonelik Erişim ve Açık Erişim dergileri arasında istatistiksel olarak anlamlı bir fark vardır ($p=0.001$) ve 2005-2008 yılları arasında da Abonelik Erişim ve Açık Erişim dergileri arasında istatistiksel olarak anlamlı bir fark vardır ($p=0.026$). Ancak, 2010-2018 ve 2015-2018 yılları arasında Abonelik Erişim ve Açık Erişim arasında istatistiksel olarak anlamlı bir ilişki bulunamamıştır.



Şekil 2: 1999-2018 döneminde Abonelik Erişimli dergilerin Etki Faktörlerinin dağılımı



Şekil 3: 1999-2018 döneminde Açık Erişimli dergilerin Etki Faktörlerinin dağılımı

Abonelik Erişim ve Açık Erişim dergileri için "**En Küçük Kareler Yöntemi**" kullanılarak ayrı regresyon analizleri yapılmıştır (Van de Geer, 2005). Grafikler, 1999-2018 döneminde Abonelik Erişim ve Açık Erişim dergilerinin uygun değerlerini temsil etmektedir. Açık Erişim günlükleri için eğim hatası parametre tahmini

0,045'tir ve standart hatası 0,019'dur. Abonelik Erişim günlükleri için eğim parametre tahmini 0,028'dir ve standart hatası 0,004'tür. Abonelik Erişim ve Açık Erişim dergilerinin eğimleri Z testleri kullanılarak test edildi (Chen, 2011).

Z testi=$(\bar{X} - \mu)/(\sigma/\sqrt{n})$ • H0 = $\beta 1 = \beta 2$ (eğimler eşittir)

• H1 = $\beta 1 \neq \beta 2$ (eğimler eşit değildir)

Her iki eğimin de p değerleri anlamlıydı (p değeri < 0,001). Açık Erişim dergilerinin eğimi, Abonelik Erişim dergilerinin eğiminden biraz daha fazladır. Grafikler, Açık Erişim dergileri için, IF rakamlarındaki artışın, etki faktörlerinde yalnızca hafif bir artışa sahip Abonelik Erişim dergilerinden daha dik olduğunu göstermektedir. Son yıllarda, Açık Erişim dergileri için eğim daha da artmıştır. Bu, Açık Erişim dergilerinin etki faktörlerinin son zamanlarda abonelik erişim dergilerinden daha hızlı arttığı anlamına gelir.

Tartışma

Bu çalışmanın bulguları bir dizi önemli noktayı göstermektedir. İlk olarak, 2016 – 2018 yılları arasında açık erişim dergileri için Etki Faktörleri, abonelik erişimli dergiler için olanları önemli sayılabilecek ölçüde geride bırakmıştır. Açık erişim için 1.64 ve abonelik erişimi için 1.23 rakamları bunun açık bir göstergesidir. t-testi sonucuna göre, yıllık dilimlerdeki verilere bakıldığında etki faktörlerinde istatistiksel olarak anlamlı bir fark olmadığını göstermiştir. Dergilerin oranı ile ilgili olarak, yukarıdaki Tablo 2, 2018'deki 199 dergiden 34'üne kıyasla, 1999'da 101'den sadece 4'ünün açık erişim olduğunu göstermektedir. Bu, önemli bir artış göstermekte ve dolayısıyla artan bir eğilime işaret etmektedir. Dergilerin açık erişime açılması. Şekil 2 ve 3'teki daha dik bir eğim, Abonelik Erişim dergilerine kıyasla Açık Erişim dergilerinin etki faktörü rakamlarında önemli bir sıçrama olduğunu göstermektedir. Bu test, yıllar içinde Açık Erişim ve Abonelik Erişim dergilerinin oranını dikkate alsa da, bu, şekil 1'deki sonuçları yansıtmaktadır.

Bu çalışmanın öne çıkan bulgusu, açık erişimli dergilerde İstatistik ve Biyoistatistik alanında yayınlanan çalışmaların kalitesinin olumsuz etkilenmemesidir. Bu nedenle, abonelik ve açık erişim dergi kalitesi arasındaki fark belki de yıllar önce şu an olduğundan daha da büyük olabilir.

Çözüm

Biyoistatistik ve istatistik alanında, akademik dergiler tarihsel olarak abonelik erişimli dergilerde yer almışlardır. Bu çalışma, açık erişim dergilerinin oranının ve dergi kalitesinin (SCImago verilerine dayalı olarak) arttığını göstermiştir. 2016'dan itibaren, bu alandaki açık erişim dergilerinin etki faktörlerinin keskin bir şekilde arttığı görülüyor; bu, belki de artan kalitelerinin ve bilimsel geçerliliklerinin önemli bir göstergesidir. Eğilim, açık erişim dergilerinin daha popüler hale gelmesi ve çalışmalarını hangi platformlardan/veri tabanlarından yayınlayacaklarını ve dağıtacaklarını seçerken araştırmacılar için avantajlı seçenekler olarak görünmektedir.

Kaynaklar

1. Garfield E. *Citation analysis as a tool in journal evaluation. Science* 1972; 178:471-9.
2. Garfield E. *The history and meaning of the journal impact factor*. JAMA 2006; 295:90-3.
3. Kazikdas K.C., Tanik M., Ural A. (2019) *Changing Trends in Otorhinolaryngology publishing*, Acta Otorhinolaryngologica Italia ,Epub : March 25
4. Polat B., Ozmanevra R., Ozmanevra P.T., Kazikdas K.C. (2019) *Comparison of the impact factors of subscription access and open access orthopedics and sports medicine journals in the SCImago Database*, Joint Diseases and Related Surgery (2019) 30 (2) 163-167
5. *SCImago Journal and Country Rank Database*: https://www.scimagojr.com/
6. Van de Geer S.A. *'Least Squares Estimation' in Everitt B.S. and Howell D.C. (Eds) (2005) Encyclopedia of Statistics in Behavioural Science*, Wiley & Sons (Chichester)
7. Van Noorden R. *Open access: The true cost of science publishing*. Nature 2013; 495:426-9.
8. Zhongxue Chen (2011) *Is the weighted z -test the best method for combining probabilities from independent tests?* Journal of Evolutionary Biology

KS10

Avrupa Sağlık Okuryazarlığı Ölçeği Türkçe versiyonu Kısa Form Önerisi

Mehmet Ali SUNGUR[1], Zerrin GAMSIZKAN[2], Demet Hanife SUNGUR[3]

[1]*Düzce Üniversitesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Düzce*
[2]*Düzce Üniversitesi, Aile Hekimliği Anabilim Dalı, Düzce*
[3]*Düzce Üniversitesi, Elektrik-Elektronik ve Bilgisayar Mühendisliği Anabilim Dalı, Düzce*

Amaç: Sağlık hizmeti sağlayıcıları ve sağlık politikası belirleyicilerin daha etkili müdahalelerde bulunabilmek ve sağlık koşullarını iyileştirebilmek adına bireylerin sağlık okuryazarlığı düzeyini değerlendirebilmeleri önemlidir. Sonuçlandırma ve uygulamaya koyma aşamalarının hızlı olabilmesi için bu değerlendirmenin kapsamlı olması kadar hızlı ve pratik bir şekilde yapılabilmesi de önemlidir. Bu çalışmanın amacı, 47 maddelik Avrupa Sağlık Okuryazarlığı Ölçeği Türkçe versiyonu (ASOY-TR) için bir kısa form önermektir.

Yöntem: Bu çalışmada kullanılan veriler, Düzce Üniversitesi'nde öğrenim gören 686 öğrenciye ASOY-TR uygulanan çalışmadan elde edilmiştir. Veri seti geliştirme ve doğrulama süreçleri için tamamen rasgele şekilde iki alt gruba bölünmüştür. Kısa formda yer alacak maddelerin seçimi için faktör yükleri, regresyon katsayıları, madde toplam korelasyonu ile dört farklı kısa form test seti oluşturulmuş ve 47 maddelik ASOY-TR toplam puan varyansının ne kadarının açıklanabileceği açısından karşılaştırılmıştır. Doğrulama sürecinde önerilen kısa formun yapı geçerliliğini test etmek için doğrulayıcı faktör analizi kullanılmış, model uyum indeksleri incelenmiştir.

Bulgular: Kısa form için oluşturulan test setleri ile 47 maddelik ASOY-TR ölçeğinden elde edilen puanlar arasında yüksek düzeyde korelasyon (0,91 ile 0,96 arasında) olduğu ve 47 maddelik ASOY-TR varyansını açıklama oranlarının 0,83 ile 0,93 arasında olduğu görülmüştür. En yüksek açıklama oranına sahip olan test seti ASOY-TR için en uygun kısa form (ASOY-TR-KF12) olarak seçilmiştir. Doğrulama için ayrılan veri setinden elde edilen puanlar incelendiğinde, ASOY-TR ile elde edilen puanın varyansının 0,92 oranında açıklanabildiği görülmüştür. Modelin $\chi 2/df=1,87$ (<3), RMSEA=0,043 (<0,05) ve RMR=0,023 (<0,05) değerleri ile iyi uyum gösterdiği sonucuna ulaşılmıştır. Ek olarak, CFI=0,953, NFI=0,906, GFI=0,969 ve AGFI=0,952 olup bu indekslerin de 1'e yakın olduğu ve bu doğrultuda modelin iyi uyum gösterdiği görülmüştür.

Sonuç: Bu çalışmada, önerilen kısa form ile elde edilen puanlar ile 47 maddelik ASOY-TR ölçeğinden elde edilen puanlar arasında yüksek düzeyde korelasyon saptanmış, 47 maddelik ASOY-TR ölçeğinden elde edilen puanın %92,4 oranında açıklanabildiği tespit edilmiştir. Doğrulayıcı faktör analizi sonuçlarına göre, modelin anlamlı olduğu ve iyi uyum gösterdiği görülmüştür. Ayrıca, önerilen kısa formun, ölçeğin tam hali için önerilen 12 alt boyuttan birer madde içeriyor olması da dikkate alınarak bu kısa formun Türkiye'de sağlık okuryazarlığının hızlı ve güvenilir bir şekilde değerlendirilmesi amacıyla yararlı bir kısa form olarak kullanılabileceği düşünülmektedir.

Anahtar Sözcükler: Sağlık okuryazarlığı, Avrupa Sağlık Okuryazarlığı Ölçeği, Türkçe versiyon, kısa form.

KS11

Nedensel Aracılık Analizi ve Hemşirelikte Bir Uygulama

Nihan ÖZEL ERÇEL[1], Gülhan TEMEL[1], Hilal ALTUNDAL[2], Mualla YILMAZ[2], Cemil ÇOLAK[3]

[1] Mersin Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Mersin
[2] Mersin Üniversitesi Tıp Fakültesi Eğitim ve Araştırma Hastanesi, Ruh Sağlığı ve Hastalıkları Hemşireliği Anabilim Dalı, Mersin
[3] İnönü Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişim Anabilim Dalı, Malatya
nihanozel88@gmail.com

Giriş: Araştırmalarda çoğunlukla üzerinde durulan bağımlı ve bağımsız değişkenlerin yanı sıra "aracı" olarak adlandırılan değişkenler de mevcuttur. Genel söylemiyle aracı değişkenler, bağımlı ve bağımsız değişkenler arasındaki nedensel ilişkinin daha derin ve daha sadeleştirilmiş hale gelmesini sağlayan üçüncül değişkenlerdir. Genellikle açıklayıcı ve sonuç değişkenleri arasındaki ilişkilerin ortaya konması ve modellenmesinde klasik regresyon yöntemlerinden yararlanılmaktadır. Ancak klasik yöntemler araştırılmak istenen ilişkinin ve bağımsız değişkenin sonuç değişkeni üzerindeki etkisinin ne kadarının doğrudan ne kadarının ise dolaylı bir şekilde aktarıldığına dair yeterli bilgi vermemektedir. Bu sebeple, ortaya çıkan bu ihtiyacı karşılamak için literatürde nedensel aracılık analizlerinden yararlanılmaktadır. Nedensel aracılık analizi yöntemi, regresyon yöntemlerine göre daha esnek bir yöntem olup klasik yöntemlerin uygulanabilmesi için gerekli olan varsayımların sağlanması koşulunu gerektirmemektedir. Sosyal bilimler ve psikoloji gibi alanlarda yaygın olarak kullanılan nedensel aracılık analizleri, klinik çalışmalar ve hemşirelik uygulamalarında da giderek artan bir öneme sahiptir.

Yöntem: Literatürde yapısal eşitlik modellemesi çerçevesinde formüle edilen aracılık analizlerinde geniş örneklem büyüklüğüyle çalışılmasının ve örneklem büyüklüğünün 300'ün altında olmamasının uygun olduğu belirtilmektedir. Bu bilgi göz önünde bulundurularak, kişisel bilgi formu ve kullanım izinleri alınmış ölçekler (Depresyon Damgalama Ölçeği (DDÖ), Psikolojik Yardım Almaya İlişkin Tutum Ölçeği Kısa Form (PYAITO-KF), BECK Depresyon Envanteri) kullanılarak 302 hemşireye ulaşılıp veriler toplanmıştır. Veri analizinin ilk aşamasında, sonuç değişkeni ile bağımsız ve aracı değişkenler arasındaki ilişkinin yönü ve büyüklüğünün tespitinde Pearson korelasyon testi kullanılmıştır. İkinci aşamada, bağımsız değişkenin, aracı değişkenler vasıtasıyla sonuç değişkeni üzerindeki doğrudan ve dolaylı etkilerinin tahmini ve modellernmesi için aracı değişkenlerin modele dahil edildiği regresyon denklemleri kurulmuştur. Bağımsız değişkenin sonuç değişkeni üzerindeki doğrudan etkisi, aracı değişkenlerin sonuç değişkeni üzerindeki dolaylı etkilerinin incelenmesi için uygulanan aracılık analizi yöntemi IBM SPSS PROCESS eklentisi kullanılarak gerçekleştirilmiştir. Her bir etki (doğrudan etki, dolaylı etki ve toplam etki) için %95 bootstrap güven aralıkları verilmiş olup, güven aralığı sıfırı içermediği durumda söz konusu etkilerin istatistiksel olarak anlamlı olduğu kabul edilmiştir.

Bulgular: Bağımsız değişken olarak ele alınan algılanan damgalama düzeyi ile birinci aracı değişken olarak modele alınan kişisel damgalama düzeyi arasında istatistiksel olarak anlamlı bir ilişki mevcut iken ikinci aracı değişken olarak modele alınan BECK depresyon düzeyi ve sonuç değişkeni olan psikolojik yardım almaya ilişkin tutum puanlarının bağımsız değişken algılanan damgalama düzeyi ile aralarında istatistiksel olarak anlamlı bir ilişkisi bulunmamıştır ($r_{algılanan\text{-}kişisel}$=0,150 p=0.009, $r_{algılanan\text{-}BECK}$=-0,035 p=0.541, $r_{algılanan\text{-}tutum}$=-0,14 p=0.812).

Bağımsız değişken olan algılanan damgalama düzeyinin sonuç değişkeni üzerindeki etkisinin ne kadarının doğrudan ne kadarının dolaylı olduğuna yönelik aracı değişkenler ile kurulan model sonuçlarına bakıldığında, algılanan damgalama düzeyinin psikolojik yardım almaya ilişkin tutum üzerindeki toplam ve doğrudan etkileri istatistiksel olarak anlamlı bulunmamıştır ($\beta_{toplam\ etki}$=-0,029; p=0.357 ve $\beta_{doğrudan\ etki}$=0,004 ve p=0.897). Aracı değişken olarak modele dahil edilen kişisel damgalamanın psikolojik yardım almaya ilişkin tutum üzerindeki aracılık etkisi yani dolaylı etkisi istatistiksel olarak anlamlı bulunmuş olup, BECK Depresyon düzeyinin psikolojik yardım almaya yönelik tutum üzerindeki aracılık etkisi istatistiksel olarak anlamlı bulunmamıştır ($\beta_{kişisel}$=-0,030, [-0,070_-0,006] ve β_{BECK}=-0,002, [-0,016_-0,005]).

Sonuç: Bağımsız değişkenin, sonuç değişkeni üzerindeki doğrudan etkisi ile bağımsız değişkenin aracı değişkenler vasıtasıyla sonuç değişkeni üzerindeki dolaylı etkilerinin incelenmesi için kurulan modellere bakıldığında; algılanan damgalamanın psikolojik yardım almaya ilişkin tutum üzerindeki toplam ve doğrudan etkileri istatistiksel olarak anlamlı bulunmamıştır. Aracı değişken olarak modele dahil edilen kişisel damgalama

ve BECK depresyon düzeyinin psikolojik yardım almaya ilişkin tutum üzerindeki toplam dolaylı etkisi istatistiksel olarak anlamlı bulunmuş olup kişisel damgalama düzeyinin psikolojik yardım almaya ilişkin tutum üzerindeki dolaylı etkisi istatistiksel olarak anlamlı, BECK Depresyon düzeyinin psikolojik yardım almaya ilişkin tutum üzerindeki dolaylı etkisi ise istatistiksel olarak anlamlı bulunmamıştır.

Anahtar Sözcükler: Aracı değişken, nedensel aracılık analizi, dolaylı etki

KS12

Popülasyon Genetiği Çalışmalarında Mantel Testi Üzerinde bir Uygulama

Osman Tolga KASKATI[1], Mustafa Murat ARAT[2], Fatma KAYMAKAMTORUNLARI DENİZ[3], Emre KESKİN[4]

[1] *Lokman Hekim Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara*
[2] *Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara*
. Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara
[4] *Ankara Üniversitesi, Ziraat Fakültesi, Su Ürünleri Mühendisliği Bölümü, Ankara*

Özet

Giriş: Popülasyon genetiği, kendi arasında eşleşebilen bireylerin oluşturduğu gruplar arasındaki ve içindeki genetik çeşitlilik modeli veya genetik yapı ile ilgili çalışmalar yapan bir genetik alt dalıdır. Farklı popülasyonlardaki bireyler arasındaki genetik farklılıkların tahmini, popülasyon genetiğindeki deneysel çalışmaların önemli bir bileşeni olmuştur. Bu çalışmanın amacı, Türkiye Denizlerinde yaşayan bir hamsi türü olan "engraulis" balığının 16 farklı bölgedeki popülasyonlarının kendi içindeki genetik farklılıkları ile coğrafi farklarının arasındaki ilişkiyi MANTEL testi ile incelemektir.

Yöntem: Popülasyonlar arasındaki genetik farklılıkların belirlenmesinde kullanılan farklı istatistiksel yöntemler mevcuttur. Özellikle genetik sapma veya genetik mesafelerin coğrafi mesafelerle karşılaştırılmak ve popülasyon yapısını etkileyen mekânsal süreçleri değerlendirmek için kullanılan yöntemlerden biri Mantel testidir. İlk olarak iki matris arasındaki ilişkiyi test etmek için önerilmiş olan Mantel testi özellikle genetik veriler için incelenen hipotezin yalnızca mesafeler cinsinden formüle edilebildiği birkaç uygun testten biridir. Mantel Testi, matrislerdeki girdilerin korelasyonunu hesaplama, matrisleri permüte etme, her permütasyon altında aynı test istatistiğini hesaplamaktan ve orijinal test istatistiğini, permütasyonlardan gelen test istatistiğinin dağılımıyla karşılaştırarak bir p-değeri üretme adımlarından oluşur.

Mantel testi pearson korelasyon katsayısının daha detaylı ve kompleks yapıda DNA dizilimleri ile oluşmuş veri kümeleri arasındaki ilişkisel yapıyı analiz etmek için konumlandırılmıştır. Bu analiz, genetik ve coğrafi mesafelerin korelasyonu üzerinden standartlaştırma yapılması ile mümkün olabilmektedir. Mantel testi, coğrafi uzaklık ve genetik sapma arasındaki ilişkiyi değerlendirmek için uzun yıllandır kullanılmaktadır. Uzamsal/mekânsal çok değişkenli verileri analiz etmek için farklı yaklaşımlar olmasına rağmen Mantel testi yaygın olarak uygulanmaktadır. Çalışmada, Ege denizinde bulunan engraulis isimli hamsi türüne ilişkin genetik veri seti üzerinde çalışmaktır. Analizler, Python program dili üzerinde gerçekleştirilmiştir Uzaklık matrislerini hesaplamak için SciPy kütüphanesi (https://www.scipy.org), Mantel Testini uygulamak için Github'daki mantel modülü (https://github.com/jwcarr/mantel) ve Scikit-Bio kütüphanesi (http://scikit-bio.org) kullanılmıştır.

Bulgular: Ege Denizinde farklı bölgelerde bulunan Engralius isimli hamsi popülasyonlarının coğrafi mesafeleri ile bu yosunların genetik mesafeleri arasındaki ilişki, Mantel testi ile analiz edilmiştir. Çalışma bünyesinde, Türkiye'deki 16 farklı istasyondan toplanmış, Engraulis türüne ait bireylerine ait olan veriler üzerinde uzaysal mesafeleri içeren uzaklık matrisi, ve verilen noktalarda ölçülen sonuçlar arasındaki mesafeleri içeren uzaklık matrisi olmak üzere iki matris oluşturulmuştur. İlk olarak bu 16 istasyonun enlem ve boylamları elde edilmiştir ve aşağıdaki uzaklık matrisi Öklid Uzaklığı kullanılarak bulunmuştur. Daha sonra, bu 16 istasyondaki bireylerden hesaplanan genetik uzaklık değişkeni kullanılarak ikinci bir uzaklık matrisi hesaplanmıştır. İki matris arasında gözlenen korelasyon, r=0.16, matris girdilerinin pozitif olarak ilişkili olduğunu göstermiştir. p değeri 0,05'ten küçük olduğundan (veya alternatif olarak, z-skoru %95 güven düzeyinde 1,96'dan büyük olduğundan), bu iki uzaklık kümesi arasında önemli bir korelasyon olduğu sonucuna varabiliriz.

Sonuç: Mantel testi uygulaması sonrasında çıkan korelasyon değeri, birbirine daha yakın yaşayan bireylerin genetik olarak daha fazla ilişkili olma eğilimindeyken, daha uzakta yaşayan bireylerin daha az genetik olarak ilişkili olma eğiliminde olduğunu göstermektedir. Karmaşık yapıda olan genetik popülasyon veri setleri ve matrislerinin karşılaştırılmasında mantel testinin kullanım ve uygulama aşamalarının benzer çalışmalarda kullanılarak coğrafi uzaklıklar ve genetik farklılıkların canlı popülasyonları üzerindeki ilişkilerinin incelenmesi için örnek bir çalışma gerçekleştirilmiştir.

Anahtar Sözcükler: Popülasyon genetiği, Genetik mesafe, Mantel testi

Kaynaklar

1. Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. D. C., ... & Bini, L. M. (2013). Mantel test in population genetics. Genetics and molecular biology, 36, 475-485.
2. Legendre, P., & FORTIN, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular ecology resources, 10(5), 831-844.
3. Legendre, P., Fortin, M. J., & Borcard, D. (2015). Should the Mantel test be used in spatial analysis?. Methods in Ecology and Evolution, 6(11), 1239-1247.
4. Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. D. C., ... & Bini, L. M. (2013). Mantel test in population genetics. Genetics and molecular biology, 36, 475-485.

KS13

What Is Synthetic Control Method And Why Should It Be Used?

Özge Pasin[1], Handan Ankaralı[2]

[1] Bezmialem University, Faculty of Medicine, Biostatistics Department, Istanbul
[2] Istanbul Medeniyet University, Faculty of Medicine, Biostatistics Department, Istanbul

Introduction: The most well-known technique used when making causal inferences is the difference-in-difference approach. In this approach, one group includes pre- and post-intervention measurements, while the other group takes measurements without any intervention. The temporal variation of the untreated group, called the control group, is an indicator of the natural trend. However, when the trends in the pre-intervention measurements of the control and study groups are not similar, the use of the synthetic control group is recommended. In this study, it was aimed to explain the theoretical features of the synthetic control method, which has not yet become widespread in the literature in terms of application, and to define its usage areas.

Method: The theoretical features of the synthetic control method are explained and its usage areas in the literature are shown. In addition, the latest developments on the subject have been analyzed and summarized. In the study, the application of the method is shown by giving sample codes for the application.

Results: When Pubmed was searched with the keyword "synthetic control method", it was found that there were 74 studies in the world literature. However, almost half of these studies belong to the years 2020 and 2021. When we evaluate it in terms of health research, it has been seen that the method is used to evaluate the health policies of countries in a significant part of the studies. In a study conducted in 2020, a review was prepared about the use of a synthetic control group for clinical trials and the difficulties experienced in assigning and/or posting individuals to the placebo group for randomized controlled studies were mentioned, and it would be a correct approach to create a synthetic control group because the patients had expectation of treatment. they have stated. Recently, this method has become valid by the FDA and has been used in studies.

Conclusion: As a result, we think that the synthetic control method, which has been used in clinical trials for the last two years in the world literature, should be used in studies in our country. We also suggest that the method should be used to evaluate the policies in health studies.

Keywords: Synthetic control, randomized, intervention

References

1. Abadie, A., Diamond, A.,Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. Journal of Statistical Software, 42(13).
2. Athey, Susan, Guido W. Imbens. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. Journal of Economic Perspectives, 31 (2).
3. Thorlund, K., Dron, L., Park, J. J., Mills, E. J. (2020). Synthetic and external controls in clinical trials–a primer for researchers. Clinical Epidemiology, 12.

KS14

Genel Özellikleri ile Hata Ağacı Analizi

Kübranur TUNÇAY[1], Prof. Dr. Sıddık KESKİN[2]

[1] Van Yüzüncü Yıl Üniversitesi, Tıp Fakültesi Biyoistatistik Anabilim Dalı, Yüksek Lisans Öğrencisi
[2] Van Yüzüncü Yıl Üniversitesi, Tıp Fakültesi Biyoistatistik Anabilim Dalı

Giriş: Hata ağacı analizi (Fault tree analysis, FTA), önceden tanımlanmış bir olayın oluşmasına neden olan paralel veya birbirini izleyen hataları belirleyen bir grafik çalışmasıdır. Hatalar, sistem hatası, insan hatası veya diğer hataların kombinasyonları olabilir. FTA, tepe olaya götüren temel olayların arasındaki mantıksal etkileşim ve bağlantıları inceler. FTA bir kantitatif (nicel) analiz yöntemi değildir. Ancak sıklıkla kantitatif analize gidilmesini sağlar. Bu çalışmanın amacı, FTA genel özellikleri ile incelenmesi, ağacın oluşturma ilkelerinin açıklanması ve bunların nitel ve nicel birer tahmin aracı olarak kullanılabilirliğinin açıklanmasıdır.

Yöntem: Hata ağacı, (Fault tree) belirli bir sistemin durumunu, çalışma sürecinde meydana gelebilecek olası arızayı, olumsuzluğu, başarısızlığı veya hatayı; nedenleri ile birlikte mantıksal olarak ilişkilendirerek, belirli bir gösterim kuralları çerçevesinde grafiksel olarak sunar. Hata ağacında genel olarak, sonuç veya çıktı niteliğinde, analiz edilen yapının veya sistemin bir başarısızlık modu bulunur. Hata ağacı analizi aşağıdaki adımları içerir: 1) Sistemin tanımlanması (sistemi oluşturan unsurlar, bunların işlevsel ilişkileri ve performans gereksinimleri). 2) Sonuç veya çıktı niteliğindeki, analiz edilecek olayın tanımı ve sınırlarının belirlenmesi. 3) Sonuç (çıktı) olayının (arızanın, hatanın veya başarısızlığın) tasarım içinde belirtilen işlevsel düzeyinde, bir veya daha fazla nedene kadar izlenmesiyle hata ağacının oluşturulması. 4) Sonuç (çıktı) olayının (arızanın, hatanın veya başarısızlığın) olasılığının veya sıklığının hesaplanması. 5) Analiz sonuçlarının raporlanması.

Bulgular: Bir hata ağacının nicel analizi, sonuç veya çıktı olayına ulaşmak için hangi olayların en kritik olduğunu ve dolayısıyla iyileştirmelerin en yararlı olacağı kısımları (bölgeleri veya yerleri) belirlemenin yollarını sağlar. Ayrıca, ana olayların meydana gelme olasılıkları verildiğinde, sonuç (çıktı) olayının meydana gelme (veya gerçekleşmeme) sıklığını veya olasılığını tahmin etmek için araçlar sağlar. Hata ağacı analizinde kantitatif analiz, 'ağacın işlenmesi' ve 'kesim setleri (kümeleri)' olmak üzere iki ana yöntemle yapılır. İlk yöntemde, her kapının (geçit) sırayla hesaplanması gerektirir. Unsurlar bağımsız olduğundan, hesaplamalar, sonuç olayına ulaşılana kadar her kapı için tekrarlanır. İkinci yöntem, sonuç olayına yol açan tüm kesme kümelerini tanımlar, her birini değerlendirir ve sonuçları toplar. Bu yöntemin bağımlı hatalarla başa çıkma veya bu hataların üstesinden gelme avantajı vardır.

Sonuç: Hata ağacı analizi (FTA) sistem tasarımını analiz etmek, değişmesi gereken parçalar üzerinde çalışmalar yapmak, 'bağımlı arızaları' analiz etmek, sistemin güvenlik gereksinimlerini ve uygunluğunu değerlendirmek ve tasarım iyileştirmesine ilişkin yapılacak eklemelerin doğruluğunu kontrol etmek gibi konularda nesnel bir temel yaklaşım sunmaktadır.

Anahtar Sözcükler: Güvenilirlik, emniyetli sistem, risk, başarısızlık, kantitatif analiz

Kaynaklar:

1. Şenel, B., Şenel, M., 2013. Risk Analizi: Türkiye'de Gerçekleşen Trafik Kazaları Üzerine Hata Ağacı Analizi Uygulaması. Anadolu Üniversitesi Sosyal Bilimler Dergisi, Cilt/Vol.: 13- Sayı/No: 3 (65-84).
2. Erdoğan, A., 2015. Hata Ağacı Analizi, Literatür Araştırması ve Orta Ölçekli Bir İşletmede Uygulama. ÇSGB Çalışma Dünyası Dergisi / Cilt: 3 / Sayı: 1 / Ocak-Nisan 2015 / Sayfa: 106-122.
3. Lee, W.S., Grosh, D.L., Tillman, F.A., Lie, C.H., 1985. Fault Tree Analysis, Methods, and Applications – A Review. IEEE Transactions on Reliability, vol. R-34, no. 3, pp. 121–123.
4. Menteş, A. & Helvacıoğlu, İ., 2010. Çok Noktadan Bağlı Tanker-Şamandıra Bağlama Sistemlerinde Hata Ağacı Tabanlı Risk Analizi. Gemi ve Deniz Teknolojisi, Sayı: 182, Şubat 2010

BİYOİSTATİSTİK DERNEĞİ 1997

ÇEVRİM İÇİ KONGRE

22. Ulusal ve 5. Uluslararası

BİYOİSTATİSTİK KONGRESİ

28-30 EKİM 2021

DÜZENLEME KURULU

Prof. Dr. Erdem KARABULUT Hacettepe Üniversitesi - Başkan
Prof. Dr. Seval KUL Gaziantep Üniversitesi
Prof. Dr. Mehtap AKÇİL OK Başkent Üniversitesi
Doç. Dr. Derya GÖKMEN Ankara Üniversitesi
Doç. Dr. İlker ÜNAL Çukurova Üniversitesi
Uzm. Ebru ÖZTÜRK Hacettepe Üniversitesi

BİLİMSEL SEKRETERYA

Prof. Dr. Seval KUL
Gaziantep Üniversitesi
Biyoistatistik Anabilim Dalı
GSM : 0 506 281 3875
E-posta : sevalkul@yahoo.com

ORGANİZASYON SEKRETERYASI

Coraltravel

Bilgi İçin : Cahit Sıtkı Bozkurt
GSM : 0 532 302 0848
E-mail : info@biyoistatistikkongresi.net
Telefon : +90 212 247 46 46
Belgegeçer : +90 212 225 72 64

www.biyoistatistikkongresi.net