ORİJİNAL MAKALE / ORIGINAL ARTICLE

Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi / DÜ Sağlık Bil Enst Derg
Journal of Duzce University Health Sciences Institute / J DU Health Sci Inst
ISSN: 2146-443X sbedergi@duzce.edu.tr 2021; 11(2): 129-136
doi: https://dx.doi.org/10.33631/duzcesbed.784688

İstatistiksel Eşleme Metodolojisi ve Rubin Eşleme Yöntemi' nin Sağlıkta Kullanımı ile İlgili Ampirik Bir Değerlendirme

İsmet DOĞAN [1], Nurhan DOĞAN [1]

ÖZ

Amaç: Bu çalışmanın amacı, istatistiksel eşleme yöntemlerini değerlendirmek ve Rubin' in istatistiksel eşleme yöntemini sağlık alanında örnek bir uygulama ile tanıtmaktır.

Gereç ve Yöntemler: İstatistiksel eşleme yapay mikro veri setleri oluşturmak için bir yöntem olarak son yıllarda giderek artan bir popülariteye sahiptir. İstatistiksel eşleme, bir araştırmada aynı anda gözlenmemiş (Y, Z) rasgele değişken çiftinden elde edilen taslak bilgi problemini ele almaktadır. Gerçekte Y ve Z birbirinden bağımsız iki farklı araştırmada birbirleri ile örtüşmeyen gözlem birimlerinin oluşturduğu kümelerden elde edilmektedir. Ancak iki araştırmada aynı X değişkeni ortaklaşa gözlenmektedir. İstatistiksel eşleme yöntemleri iki farklı veri kümesinden elde edilen bilginin birleştirilmesini hedeflemektedir.

Bulgular: Eşleme işleminde hangi veri setinin alıcı hangisinin donör veri seti olacağı ve kohort değişken kullanmanın söz konusu olup olmayacağı önem arz etmektedir. Çünkü bunlar hem eşlemede hem de eşleme sonucunda hesaplanan uzaklık ölçüsünün değerinin belirlenmesinde belirleyici olmaktadır. Özellikle kohort değişken kullanılması uzaklık ölçüsünün değerini minimum olmaktan uzaklaştırmaktadır. Literatürde eşleme ile ilgili yöntemler tam eşleme, kısıtlı ve kısıtsız istatistiksel eşleme isimleri ile yer almaktadır. Bu isimlendirmelere tam ve istatistiksel eşleme yöntemlerinin karışımı olan kısıtlı ve kısıtsız aşamalı eşleme isimleri de ilave edilebilir.

Sonuç: Rubin tarafından önerilen yöntem, diğer yaklaşımlara göre oldukça iyi sonuçlar vermesine rağmen en iyi yöntem veya yöntemler konusunda fikir birliği bulunmamaktadır. En iyi yöntem veya yöntemlere ilişkin görüş birliği bulunmadığından kısıtlı ve kısıtsız yöntemler halen kullanılmaktadır.

Anahtar Kelimeler: Dosya birleştirme; istatistiksel eşleme; Rubin istatistiksel eşleme yöntemi; veri birleştirme; veri karıştırma; yapay eşleme.

Statistical Matching Methodology and An Empirical Evaluation of the Use of Rubin's Matching Method in Health

ABSTRACT

Aim: The aim of this study is to evaluate statistical matching methods and introduce Rubin's statistical matching method with an exemplary application in the field of health.

Material and Methods: Statistical matching has received increasing popularity in the last decades as a method of creating synthetic microdata sets. Statistical matching tackles the problem of drawing information on a pair of random variables (Y, Z) which have not been observed jointly in one sample survey. In fact, Z and Y are available in two distinct and independent surveys whose sets of observed units are non overlapping. The two surveys observe also same common variables X. Statistical matching techniques are aimed at combining information available in two distinct datasets.

Results: In the matching process, it is important that which data set will be the recipient and which will be the donor data set and whether it is possible to use cohort variables. Because these are decisive both in matching and determining the value of the distance measure calculated as a result of the matching. In particular, the use of cohort variables makes the value of the distance measure away from being minimum. In the literature, methods related to matching are included with the names of exact matching, constrained and unconstrained statistical matching. Restricted and unconstrained progressive matching names, which are a mixture of exact and statistical matching methods, can also be added to these

1 Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Afyon, Türkiye

Sorumlu Yazar / Corresponding Author: Nurhan DOĞAN, e-mail: nurhandogan@hotmail.com
Geliş Tarihi / Received: 16.10.2020, Kabul Tarihi / Accepted: 21.01.2021

nomenclatures.
Conclusion: Although the method proposed by Rubin gives very good results compared to other approaches, there is no consensus on the best method or methods. No consensus regarding the best method or methods has developed; both constrained and unconstrained methods are still being used.
Keywords: File concatenation; statistical matching; Rubin's statistical matching method; data merging; data fusion; synthetical matching.

GİRİŞ

İstatistiksel eşleme ampirik çalışmalarda yaygın olarak kullanılan bir tekniktir ve tek bir veri setinin önemli çıkarımlar sağlamak için gerekli tüm bilgileri içermediği durumlarda uygulanmaktadır. Bu yöntem, iki bağımsız veri setinde yer alan bilgilerin istatistiksel çıkarımların elde edilebileceği tek bir birleşik veri seti olarak birleştirilmesini içermektedir. İstatistiksel eşlemede amaç, her iki veri setinin de aynı popülasyonu temsil ettiği varsayımı altında veri setlerinde bulunan ortak bilgileri kullanarak veri setlerinin birleştirilmesini sağlamaktır (1). Önemli çıkarımlar elde edebilmek amacıyla birçok durumda, ortak değişkenler içeren birkaç mikro veri setinden araştırmanın içeriğine uygun tek bir mikro veri seti oluşturmak gerekebilir. Tek bir dosya gerekli değişkenlerin tam setine sahip olmadığında, ortaya çıkan pratik sorunu çözmek için istatistiksel olarak kayıtlar eşleştirilir (2). Kayıt eşlemenin arkasındaki temel fikir, A ve B gibi iki orijinal dosyadaki tüm veri öğelerini içeren birleşik bir C dosyası oluşturmak için A ve B dosyalarındaki kayıtları birleştirmektir. Bu, her iki veri seti için ortak olan değişkenlere göre eşlenecek kayıt çiftlerini seçerek gerçekleştirilir. İstatistiksel eşleme (veri füzyonu, veri birleştirme, dosya birleştirme veya sentetik eşleme olarak da adlandırılır), birkaç ortak öğe kullanarak benzer kayıtları birleştirir. Benzer kayıtların eşleştirilmesi ile oluşturulan C dosyası iki farklı bireyin/birimin birleşimi olabilen ancak öznitelikleri araştırma amaçları için yeterince benzer olan kayıtları içerir (3). Tek bir veri setinde bulunan değişkenlerin gerekli analiz için yetersiz olduğu birçok durum vardır. Yeni verilerin toplanması çok pahalı veya imkansız ise, araştırmacı kayıtlarını mevcut veri setlerindeki verilerle ilişkilendirerek ek bilgiler elde etmeye çalışabilir. Böyle bir bağlantı ya tam eşleme ya da istatistiksel eşleme yöntemleri kullanılarak gerçekleştirilebilir. Tam eşlemede aynı bireye/birime ait olan değişkenler birleştirilir. Bu yöntem bireye/birime ait özel tanımlayıcılar (isim, adres, kimlik numarası vb.) gerektirir. Aynı olmayan bireyin/birimin eşlenmesi sadece hata sonucu olarak ortaya çıkar (4). Tam eşlemede, ilişkili kayıtların eşleşen aday olarak kabul edilmesi için eşlenen değişkenlere ait bilgilerin hatasız olması gerektiğini belirtmek gerekir. Bu durumun "doğruluk" ya da "hatasız" eşlemelerle ilgisi yoktur. Bunun nedeni, eşleştirilecek veri setlerinde ve daha özel olarak eşlenen değişkenlerde hatalar, sapmalar veya düzensizliklerin ortaya çıkmasıdır. Verilerdeki bu hatalar, aynı birimlerle ilgili olmayan eşlenen adayların bulunduğu bir duruma yol açar (5). Aynı bireyleri/birimleri içeren veri setlerinin eşleştirilmesi durumu dikkate alındığında, en azından sürekli ve ölçüm hataları içermeyen ortak bir değişken etkili bir tanımlayıcı olarak kullanılabilir. Çünkü söz konusu değişkene ait birliktelik olasılığı diğer bireyler/birimler için sıfırdır. Bu nedenle, bu özel durumda iki veri tabanını bir araya getirmek için istatistiksel eşleşmeye gerek yoktur. Ayrıca eşlemede dikkate alınan veri tabanlarındaki örtüşen değişkenlerin kesikli veya kategorik değişken olması durumunda, bunlar kohort değişkenler (eşleme için koşul olarak belirlenen değişken) olarak işlev görür ve kohort değişkende hiçbir değişiklik olmadan alt veri tabanları oluşturulabilir. Örtüşen değişkenlerin bazıları kesikli bazıları da hatalı ölçüm değerlerine sahip sürekli değişkenler ise eşleştirilen bireyler/birimler arasındaki farklılıkları dağıtmak için kayıt paketleri oluşturulur. Kayıt paketleri, eşleştirilen birimler arasındaki farklılıkları yaymak için oluşturulan kayıtlardır. Paket düzeyinde eşleşen kayıtların seçiminde, eşlemelerde kullanılan uzaklık ölçülerinden yararlanılır. En küçük uzaklığa sahip eşleşme, eşleme olarak seçilir (6). Uzaklık birçok şekilde tanımlanabilir. Öklid uzaklığı, City-Block uzaklığı ve Mahalanobis uzaklığı en sık kullanılan uzaklık metrikleridir (7). Eşleme yapılırken alternatif çözümler arasından seçim yapılmaktadır. Seçim işlemi donör veri setinde yer alan her bir kaydın alıcı veri setinde yer alan her bir kayıt ile eşleşme alternatifleri ile karakterize edilmektedir. Eşlemede büyük önem taşıyan değişkenler ve ilişkiler açısından donör ile alıcının mümkün olduğu kadar birbirine benzer olduğu bir çözüm seçilmelidir. Bu yaklaşım bir "uzaklık fonksiyonu" sorunu olarak görülebilir. Genel anlamda donör ile alıcı arasındaki uzaklığı ölçen bir uzaklık fonksiyonu tanımlanabilir. Uzaklığı en aza indiren istatistiksel eşleşme, en uygun eşleşme sonucudur (8). İstatistiksel eşleme problemi, bazı değişkenlerin birlikte gözlenmediği çoklu veri kümelerinin entegrasyonunu içerir (9). İstatistiksel eşleme benzer olan kayıtları bir araya getirmektedir, kayıtların aynı bireye/birime ait olmasına gerek yoktur. Yani her bir birey/birim veri setlerinden yalnızca birinde gözlenebildiği gibi diğer veri setinde de gözlenebilir. Veri setlerinde gözlenen birimlerin farklı olması gerekliliği gibi bir şey söz konusu değildir (10). İstatistiksel eşlemede, özdeş birimlerden ziyade benzer birimlerin eşleştirilmesi sadece kabul edilebilir değil aynı zamanda beklenmektedir. Bu nedenle istatistiksel eşleme, temel istatistiksel kaynakların çok az ortak kayda sahip olduğu veya hiç bulunmadığı durumlarda kullanılır ve bu da verilerin çoğunda aynı kayıtların eşleştirilmesini imkansız hale getirmektedir. İstatistiksel eşleşmeler, tam ve olasılıksal eşleşmelerde olduğu gibi özelliklerin benzerliği temelinde yapılır (11). İstatistiksel eşleme 50 yılı aşkın bir süredir yaygın olarak kullanılmaktadır. Mevcut prosedürler tipik ancak genellikle örtük olarak bir koşullu bağımsızlık ilişkisinin bulunduğunu varsaymaktadır. İstatistiksel eşleşme her zaman özel bir yere sahip olsa da konu ile ilgili bazı kısımlar dikkatle incelenmiştir (12). 1980'lerin başından beri, sadece sosyoloji ve iktisatta değil, aynı zamanda diğer disiplinlerde (ekonomi, epidemiyoloji, tıp ve siyaset bilimi) istatistiksel eşleşmenin popülerliği sürekli artmıştır. İstatistiksel eşleme uygulamalarında araştırmacılar, bir şeye maruz kalan bir birey/birimin maruz kalmama durumunda nasıl davranacağını bilememek gibi temel sorunla karşı karşıyadır.

İstatistiksel eşleme, bir dizi ortak arka plan özelliğini paylaşan ve bir şeye maruz kalan veya kalmayan bireyler/birimler arasındaki sonuç farkını karşılaştırarak bu kısıtlamanın üstesinden gelmeye yardımcı olur. İstatistiksel eşleme, gerekli veriler tek bir veri setinde bir arada mevcut olmadığında ya da tek bir tanımlayıcı üzerinden birbirleri ile ilişkilendirilemeyen birkaç veri kümesinin olması durumunda devreye girer. Bu sorun, yeni veri toplamanın çok maliyetli olması veya toplanan verilerin güvenilir olmaması durumunda ortaya çıkar. Bu durumlarda, istatistiksel eşleme, bir veri kümesinden ikinci kaynaktan gerekli bilgileri içeren kayıtları bağlar. Bağlantılı bilgiler aynı kişilerden değil, her iki veri setinde de aynı (veya neredeyse aynı) arka plan özniteliklerine sahip gözlemlerden gelir (13). Mikro veri analizleri genellikle tek bir kaynaktan elde edilemeyen ancak bir dizi kaynaktan elde edilebilen birimlerden veri gerektirir. İstatistiksel eşleme, bu amacı gerçekleştirmek için kullanılan bir yöntemdir. İstatistiksel olarak eşleşen veri kümelerine dayanan bulguların geçerliliği, her girdi dosyasına özgü değişkenler arasındaki ilişkiler hakkındaki temel varsayımların doğruluğuna bağlıdır (14). Farklı veri setleri için spesifik değişkenlerin yanı sıra, her iki veri setinde de gözlemlenen ve eşlemenin yapılabileceği ortak değişkenlerin olması kaçınılmazdır. Mevcut bazı istatistiksel eşleme yaklaşımları, ortak değişkenler verildiğinde spesifik değişkenlerin koşullu bağımsızlığı varsayımına dayanmaktadır (15). Koşullu bağımsızlık varsayımına göre, eşleme yapıldıktan sonra her iki veri setinde de gözlemlenen değişkenler göz önüne alındığında, eşleştirilen veri setlerinde eşzamanlı olarak gözlemlenmeyen değişkenler bağımsızdır (16). Özellikle, X' e bağlı Y ve Z arasındaki ilişki ölçüleri tahmin edilemez ve genellikle "sıfır" olduğu varsayılır. Koşullu bağımsızlık varsayımı olarak isimlendirilen bu varsayım, eşleşmeye dayalı tahminlerin kalitesini değerlendirmek için bir referans noktasıdır. Eşleme yöntemleri bu koşul geçerli olduğunda, birden çok veri setinden toplanan değişkenlerin gerçek ortak dağılımını yansıtan doğru tahminler üretecektir. Ne yazık ki, bu varsayım pratikte nadiren geçerlidir ve veri setlerinden test edilemez. Koşullu bağımsızlığın olmaması ve ek bilgi bulunmaması durumunda, modelde tanımlama sorunları olacak ve üretilen yapay veri setleri yanlış çıkarımlara yol açabilecektir (7). Koşullu bağımsızlık, varyans-kovaryans matrisi kullanılarak

$$\sigma_{YZ} - [\sigma_{YX_A} \quad \sigma_{YX_B}] \begin{bmatrix} \sigma_{X_A}^2 & \sigma_{X_A X_B} \\ \sigma_{X_A X_B} & \sigma_{X_B}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{X_A Z} \\ \sigma_{X_B Z} \end{bmatrix} = 0$$

şeklinde ifade edilmektedir (17). Eşlemede temel gereklilik, her iki veri kümesinin de aynı hedef popülasyonu tanımlamasıdır. Eşleme kalitesi büyük ölçüde,

- hedef popülasyonun doğru tanımı,
- verilerin uyumu,
- eşleşen değişkenlerin seçimi,
- uygun bir eşleme yönteminin seçimi,

gibi faktörlere bağlıdır. Ayrıca alıcı ve donör veri setlerinin her ikisinde ortak olan değişkenlerin miktarı, kalitesi ve bu değişkenlerin ortak olmayan değişkenler ile ne kadar ilişkili olduğu yüksek kaliteli bir eşleşme için çok önemlidir (18). Literatürde, eşlemede kullanılacak değişkenleri seçmek için iki ana kriter vurgulanmaktadır. İlk olarak, istatistiksel içeriklerinde hem homojenlik hem de değişkenlerin iki veri setindeki dağılımında benzerlik olmalıdır. İkinci olarak, değişkenlerin hedef değişkenlerdeki farklılıkları açıklamada anlamlı olması gerekir (19). İstatistiksel eşlemenin başarılı olması için farklı veri kaynaklarındaki bireylere/birimlere ait ortak X özelliği ile birlikte dikkate alınan diğer değişkenlerin benzer şekilde ölçülmesi gerekir. İstatistiksel eşlemenin ana fikri ortak değişkenleri kullanmak olduğundan eşleşmenin gerçekleştirilmesinde, kullanılan veri setleri arasındaki yapısal farklılıklar problem olabilir. Örneğin, eşleşen değişkenler iki veri kaynağında farklı hassasiyet seviyelerinde ölçüldüğünde, son derece sorunlu olabilir. En önemlisi, ölçümlerde yanlılık olmamalı veya yanlılık varsa, eşlemede kullanılan veri setlerinin benzer şekilde etkilenmesi gerekir. Eşlemede kullanılan ortak değişkenin veri setlerinden birinde diğer veri setine göre daha farklı yanlılığa sahip olması durumunda eşleme ortak değişkenin yanlış değerini temel alacaktır (20). Gözlemsel araştırmalarda önemli bir husus, çalışmada dikkate alınmayan ancak bağımlı ve bağımsız değişkenleri etkileyebilecek karıştırıcı faktörlerin etkisini kontrol etmektir. Eğer karıştırıcı değişkenler kontrol edilmezse sonuçlar, söz konusu bağımsız değişkenin incelenen bağımlı değişkenle gerçek ilişkisini yansıtmayacaktır. Karıştırıcı değişkenlerin etkilerini kontrol etmenin bir yöntemi, eşleştirilmiş bir vaka kontrol tasarımı kullanmaktır. Eşleştirilmiş vaka-kontrol çalışmalarında, bilinen karıştırıcı değişkenlerin etkisi, vaka denekleri ile kontrol deneklerinin bir veya daha fazla karıştırıcı değişken üzerinde eşleştirilerek kontrol edilir. Vaka-kontrol çalışmalarında eşleşen değişkenleri kontrol ederken hassasiyeti artırarak çalışma verimliliğini artırmak veya ölçülemeyen değişkenlerin analizinde kontrolü sağlamak için yaş ve cinsiyet gibi değişkenlerle eşleşme yaygın olarak kullanılmaktadır (21). Eşleme, vaka-kontrol çalışmalarında sıklıkla kullanılmasına rağmen kohort çalışmaları ile epidemiyolojik çalışmalarda da kullanılmaktadır. Kontrolleri belirlenen vakalarla eşleme yöntemi, kardiyovasküler hastalıklar, kanserler, pediatri, gastroenteroloji ve cerrahi gibi çok çeşitli epidemiyolojik çalışmalarda yaygındır (22). Bu çalışmanın amacı istatistiksel eşleme yöntemlerini değerlendirmek ve Rubin' in istatistiksel eşleme yöntemini sağlık alanında örnek bir uygulama ile tanıtmaktır. Makalede, Helsinki Deklarasyonu Prensipleri dikkate alınmıştır.

GEREÇ VE YÖNTEMLER

İstatistiksel eşleme problemi, aynı popülasyonda yer alan ve sadece kısmi örtüşmenin söz konusu olduğu farklı bireylerden/birimlerden toplanan bilgilerin birleştirilmesini içermektedir. Bazı değişkenleri ortak bazı değişkenleri ise ortak olmayan A ve B ile ifade edilen iki farklı veri seti olsun. İstatistiksel eşlemenin amacı, (X, Y) değişkenlerine sahip A veri seti ile (X, Z) değişkenlerine sahip B veri setini birleştirerek (X, Y, Z) değişkenlerinin tümünü içeren tek bir veri seti elde etmektir. Yapılan işlem Şekil 1'de gösterilmiştir.

$Veri\ Seti\ A\ (Donör)$	$Veri\ Seti\ B\ (Alıcı)$	$Yapay\ Veri\ Seti$
X, Y		
	X, Z	X, Y, Z

Şekil 1. İstatistiksel eşleme

Literatürde yer alan istatistiksel eşleme yöntemlerinin tamamı, iki farklı veri setindeki ortak değişkeni (X), tüm değişkenleri (X, Y, Z) içeren sentetik kayıtlar oluşturmak için bir köprü olarak kullanmaktadır. Daha az bilgilendirici ancak daha düşük maliyetli veri toplama veya veri çoğaltma yöntemi olan istatistiksel eşleme, yeniden doğru veri toplamadan kaynaklanan zayıflığın üstesinden gelmek için olası bir çözüm olarak önerilmiştir (23). Veri setlerini eşlemek dikkat gerektirir. Çünkü eşleşen veri setindeki öğelerin doğal bir anlamı yoktur. A veri setindeki bireyler/birimler için Z değişkeni ve B veri setindeki bireyler/birimler için Y değişkeni hakkındaki bilgiler eksiktir. Mekanik bir eşleme yöntemi bu bağlamda cazip olabilir, çünkü aynı bireylere/birimlere atıfta bulunmuş gibi X, Y ve Z değerlerini içeren bir veri seti üretecektir. Ancak, daha önce bulunmayan bilgilerin eşleme süreci tarafından yaratılamayacağı açıktır (17). Literatürde güçlü ve zayıf yönleri olan çeşitli eşleme yöntemleri mevcuttur. Belirli bir çalışma için hangi yöntemin en uygun olduğunu önceden belirlemek zordur. Genel olarak, istatistiksel eşleme için kullanılan yöntemler "kısıtlı" ve "kısıtsız" olmak üzere iki genel kategoriye ayrılmaktadır. Kısıtlı istatistiksel eşleme, iki veri setindeki tüm kayıtların kullanılmasını gerektirir ve temel olarak veri setlerinde örtüşmeyen değişkenlere ait marjinal dağılımlar korunmaktadır. Kısıtsız eşlemenin bu gereksinimi yoktur (24). Kısıtsız eşlemede her donör kaydı birden çok kez kullanılabilirken kısıtlı eşlemede her donör kaydı yalnızca bir kez kullanılabilmektedir. Kısıtlılık, donör kayıt sayısının alıcı kayıt sayısına eşit veya daha fazla olmasını gerektirmektedir (25). Birden çok kaynaktan gelen verilerin entegrasyonu için kullanılan istatistiksel yöntemler tam eşleme teknikleri ve istatistiksel eşleme teknikleri biçiminde de sınıflandırılmaktadır. Hem tam hem de istatistiksel eşleme teknikleri, makul ve istatistiksel olarak kontrol edilebilir varsayımlar altında entegre kayıtlar oluşturulmasına izin vermektedir (26).

Rubin Eşleme Yöntemi

Rubin (2) tarafından önerilen istatistiksel eşleme yönteminin, gerçek ve simüle edilmiş veri setleri kullanılarak yapılan çalışmalar ile, geleneksel yöntemlerden farklı olarak eksik değerleri yeterince doğru bir şekilde belirlediği ve böylece emsal değerlerle gerçekleştirilen ampirik analizlerin yansız tahminler verdiği gösterilmiştir. Yöntem, herhangi bir birey/birim için eşzamanlı olarak gözlemlenmeyen değişkenler arasındaki kısmi korelasyonu dikkate almakta ve başka bir veri setindeki gözlemler kullanılarak gözlemlenmeyen bir değişkenin eşleştirilmesine olanak sağlamaktadır (27). Kısıtlı eşlemeden çok daha az hesaplama çabası gerektirmesinden dolayı kısıtsız eşleme yöntemi Rubin tarafından önerilmiştir (2). Genel olarak Rubin eşlemeye, istatistiksel eşlemede kullanılacak değişkenlerin tahminlerini elde etmek amacıyla regresyon analizi kullanarak başlamayı önermektedir. Dolayısıyla Rubin yöntemi

A veri seti için,

$\hat{Y} = a_0 + aX + \varepsilon$

B veri seti için,

$\hat{Z} = b_0 + bX + \varepsilon$

regresyon denklemlerinin oluşturulması ile başlar.

B veri setinde eksik Y değişkenine ait her bir birey/birim, Y değişkeni için tahmin edilen en yakın birey/birim ile, A veri setinde eksik Z değişkenine ait her bir birey/birim Z değişkeni için tahmin edilen en yakın birey/birim ile eşleştirilir. Eşlemelerde minimum uzaklık ya da en yakın komşuluk yaklaşımı dikkate alınmaktadır. Rubin (2) tarafından önerilen yöntem kısıtsız eşleme yöntemi olarak sınıflandırılmaktadır (16). Yöntem, aşağıda verilen örnek veri setleri (A veri seti alıcı, B veri seti ise donör veri seti olarak dikkate alınmıştır) kullanılarak tanıtılmıştır. Veri setlerinde yer alan cinsiyet, yaş, total kolesterol ve trigliserid değişkenlerine ait değerler bir hastanenin Kardiyoloji polikliniğine başvuran hastalara ait kayıtlardan arşiv taraması yolu ile elde edilmiş gerçek verilerdir.

Tablo 1a. A veri seti

Birey	Cinsiyet	Yaş (X)	Total Kolesterol (Y)
A1	Kadın	23	102,5
A2	Erkek	25	139,9
A3	Kadın	28	151,2
A4	Erkek	35	118,4
A5	Erkek	42	161,3
A6	Kadın	53	158,5
A7	Erkek	55	136,8
A8	Kadın	64	170,7

Tablo 1b. B veri seti

Birey	Cinsiyet	Yaş (X)	Trigliserid (Z)
B1	Erkek	28	113,5
B2	Kadın	32	54,4
B3	Erkek	41	282,5
B4	Kadın	46	85,0
B5	Erkek	52	148,7
B6	Kadın	58	149,4

Adım 1, A veri setinden yararlanarak regresyon denklemi, $\hat{Y} = 103{,}788 + 0{,}951X$ olarak, B veri setinden yararlanarak regresyon denklemi ise $\hat{Z} = 67{,}575 + 1{,}666X$ olarak elde edilir.

Adım 2, $\hat{Y} = 103{,}788 + 0{,}951X$ denklemi kullanılarak hem A hem de B veri setlerinde yer alan X değişkenine ait gözlem değerleri için $\hat{Y}_A$ ve $\hat{Y}_{Bi}$ tahmini değerleri elde edilir. Bu değerler kullanılarak $\left|\hat{Y}_A - \hat{Y}_{Bi}\right|$ farkları bulunur.

Tablo 2. $\left|\hat{Y}_A - \hat{Y}_{Bi}\right|$ değerleri

Birey	$\hat{Y}_A$	$\left\|\hat{Y}_A - \hat{Y}_{B1}\right\|$	$\left\|\hat{Y}_A - \hat{Y}_{B2}\right\|$	$\left\|\hat{Y}_A - \hat{Y}_{B3}\right\|$	$\left\|\hat{Y}_A - \hat{Y}_{B4}\right\|$	$\left\|\hat{Y}_A - \hat{Y}_{B5}\right\|$	$\left\|\hat{Y}_A - \hat{Y}_{B6}\right\|$
A1	125,661	4,755	8,559	17,118	21,873	27,579	33,285
A2	127,563	2,853	6,657	15,216	19,971	25,677	31,383
A3	130,416	0	3,804	12,363	17,118	22,824	28,53
A4	137,073	6,657	2,853	5,706	10,461	16,167	21,873
A5	143,73	13,314	9,51	0,951	3,804	9,51	15,216
A6	154,191	23,775	19,971	11,412	6,657	0,951	4,755
A7	156,093	25,677	21,873	13,314	8,559	2,853	2,853
A8	164,652	34,236	30,432	21,873	17,118	11,412	5,706

Adım 3, Eşleme için A veri seti alıcı, B veri seti donör ve cinsiyet değişkeni kohort değişken olarak dikkate alınarak A veri setinde yer alan her bir birey/birim için ikinci adımda elde edilen $\left|\hat{Y}_A - \hat{Y}_{Bi}\right|$ farkları içerisinden en küçük olan seçilir.

Tablo 3. A veri seti alıcı, B veri seti donör olduğunda birleştirilmiş veri seti

A Veri Seti	B Veri Seti	X	Y	Z
A1	B2	23	102,5	54,4
A2	B1	25	139,9	113,5
A3	B2	28	151,2	54,4
A4	B3	35	118,4	282,5
A5	B3	42	161,3	282,5
A6	B6	53	158,5	149,4
A7	B5	55	136,8	148,7
A8	B6	64	170,7	149,4

Dikkate alınan örnekte A veri seti alıcı, B veri seti ise donör veri seti ve cinsiyet değişkeni kohort değişken olarak kullanılmıştır. Bu durumun tam tersi (A veri seti donör, B veri seti ise alıcı) durumun dikkate alınması durumunda $\hat{Z} = 67{,}575 + 1{,}666X$ denklemi kullanılarak gerçekleşen eşleşme Tablo 4a-b'de verilmiştir.

Tablo 4a. $\left|\hat{Y}_B - \hat{Y}_{Ai}\right|$ değerleri

Birey	$\hat{Y}_B$	$\left\|\hat{Y}_B - \hat{Y}_{A1}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A2}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A3}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A4}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A5}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A6}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A7}\right\|$	$\left\|\hat{Y}_B - \hat{Y}_{A8}\right\|$
B1	114,223	8,33	4,998	0	11,662	23,324	41,65	44,982	59,976
B2	120,887	14,994	11,662	6,664	4,998	16,66	34,986	38,318	53,312
B3	135,881	29,988	26,656	21,658	9,996	1,666	19,992	23,324	38,318
B4	144,211	38,318	34,986	29,988	18,326	6,664	11,662	14,994	29,988
B5	154,207	48,314	44,982	39,984	28,322	16,66	1,666	4,998	19,992
B6	164,203	58,31	54,978	49,98	38,318	26,656	8,33	4,998	9,996

Tablo 4b. B veri seti alıcı, A veri seti donör olduğunda birleştirilmiş veri seti

B Veri Seti	A Veri Seti	X	Y	Z
B1	A2	28	139,9	113,5
B2	A3	32	151,2	54,4
B3	A5	41	161,3	282,5
B4	A6	46	158,5	85,0
B5	A7	52	136,8	148,7
B6	A6	53	158,5	149,4

BULGULAR

Eşleme işleminde hangi veri setinin alıcı hangisinin donör veri seti olacağı ve kohort değişken kullanmanın söz konusu olup olmayacağı önem arz etmektedir. Çünkü bunlar hem eşlemede hem de eşleme sonucunda hesaplanan uzaklık ölçüsünün değerinin belirlenmesinde belirleyici olmaktadır. Özellikle kohort değişken kullanılması uzaklık ölçüsünün değerini minimum olmaktan uzaklaştırmaktadır. Dolayısıyla araştırmacı en başta bunlara karar vermelidir. Tablo 3 ve Tablo 4b

birlikte dikkate alındığında eşlemelerin neredeyse tıpa tıp aynı olduğu görülmektedir. Yaş değişkeni modellerde bağımsız değişken olarak dikkate alındığı için aynı cinsiyette en yakın yaşlara sahip bireyler arasında eşleme işlemi gerçekleşmektedir. Dolayısıyla bu durum, çalışmada verilen yaş değişkenine ait değerlerin her birinin veri setlerinde sadece bir kez yer almasından kaynaklanmaktadır. Eğer aynı yaşa sahip bireylerin sayısı birden fazla olsaydı eşleme farklı şekilde gerçekleşebilirdi. Literatürde eşleme ile ilgili yöntemler tam eşleme, kısıtlı ve kısıtsız istatistiksel eşleme isimleri ile yer almaktadır. Bu isimlendirmelere tam ve istatistiksel eşleme yöntemlerinin karışımı olan kısıtlı ve kısıtsız aşamalı eşleme isimleri de ilave edilebilir. Dolayısıyla eşleme ile ilgili yöntemler;

1. Tam eşleşme: Alıcı ve donör veri setlerinde yer alan bireyler/birimler birbirinin aynı ve aynı bireyler/birimler birbiri ile eşleşmektedir.
2. Kısıtlı istatistiksel eşleme: Alıcı ve donör veri setlerinde yer alan bireyler/birimler birbirinden farklı ve donör veri setinde yer alan bireyler/birimler alıcı veri setinde yer alan bireylerden/birimlerden yalnızca biriyle eşleşmektedir.
3. Kısıtsız istatistiksel eşleme: Alıcı ve donör veri setlerinde yer alan bireyler/birimler birbirinden farklı ve donör veri setinde yer alan bireyler/birimler alıcı veri setinde yer alan bireylerden/birimlerden birden fazlası ile eşleşmektedir.
4. Kısıtlı aşamalı eşleme: Alıcı ve donör veri setlerinde yer alan bireylerin/birimlerin bir kısmı aynı bir kısmı ise farklı olabilir. Birinci aşamada öncelikle alıcı ve donör veri setlerinde yer alan aynı bireyler/birimler birbiri ile eşleştirilir. İkinci aşamada ise birbirinden farklı bireyler/birimler arasında donör veri setinde yer alan bireyler/birimler alıcı veri setinde yer alan bireylerden/birimlerden yalnızca biriyle eşleşmektedir.
5. Kısıtsız aşamalı eşleme: Alıcı ve donör veri setlerinde yer alan bireylerin/birimlerin bir kısmı aynı bir kısmı ise farklı olabilir. Birinci aşamada öncelikle alıcı ve donör veri setlerinde yer alan aynı bireyler/birimler birbiri ile eşleştirilir. İkinci aşamada ise donör veri setinde yer alan ve birinci aşamada eşleşen bireyler/birimler de kullanılarak donör veri setinde yer alan bireyler/birimler alıcı veri setinde yer alan bireylerden/birimlerden birden fazlası ile eşleşmektedir.

biçiminde sınıflandırılabilir.

TARTIŞMA

İstatistiksel eşleme, aynı popülasyondan iki veya daha fazla veri seti aracılığıyla toplanan değişkenlerin ve göstergelerin ortak dağılımı hakkında bilgi sağlamak için kullanılan istatistiksel bir yaklaşımdır. Maliyetleri ve yanıt yükünü artırmadan mevcut verilerin değerini artırma imkanı sunar. İstatistiksel eşleme teknikleri, mevcut verilerdeki kondensite kısıtlamaları nedeniyle, bireysel kayıtların tam olarak eşleştirilmesinin mümkün olmadığı durumlarda, aynı hedef popülasyona atıfta bulunan iki veya daha fazla veri setini entegre etmeyi amaçlamaktadır. İki veri seti bir veya daha fazla değişkeni paylaşmak zorundadır. Temel amaç, alıcı olarak seçilen veri setini, yalnızca diğer veri setinde (donör) mevcut olan değişkenlerin değerleriyle doldurmaktır. Empütasyon yaklaşımları, eksik verileri telafi etmek için tipik bir araştırma sonrası stratejisi olarak da kullanılır. Alıcı veri setindeki her birim için, temel amaç donör veri setinde benzer varlıkları aramak ve yalnızca donör veri setinde bulunan değişkenlerin değerini belirtmektir. İstatistiksel eşleme, iki ayrı veri setinde mümkün olandan daha esnek bir analize olanak tanıyan yeni bir sentetik veri seti oluşturmak amacıyla bağlanabilecek benzer kayıtları tanımlamak için her iki veri setinde ortak olan değişkenleri kullanır. Bu amaçla, alıcı ve donör verilerinin ne kadar benzer olduğunu hesaplamak ve / veya kümelenmeyi benzerliğe dayalı bir segmentasyon sağlamak için bir uzaklık fonksiyonunun tanımlanması gerekir (28). Eşleme güçlü bir istatistiksel araçtır, ancak sihirli bir değnek değildir. Alternatif yöntemlere göre eşlemenin güçlü ve zayıf yönleri dikkatle ele alınmalı ve faydaları en üst düzeye çıkarmak için kullanımı optimize edilmelidir. Herhangi bir istatistiksel analizde olduğu gibi, eşleme çalışmaları da dikkatli bir tasarım gerektirir. Eşleme analizi için üç ana adım söz konusudur. İlk adım, alıcı ve donör veri setlerinde yer alacak bireyleri/birimleri tanımlamaktır. İkinci adım, uygun değişkenlerin ve spesifik eşleme yaklaşımının seçilmesinden oluşur. Üçüncü adım, eşleme analizini yürütmeyi ve eşlemenin kalitesini değerlendirmeyi içerir. Adım 2 ve 3, eşleme optimize edilene kadar tekrarlanır. Bu aşamalar tamamlandıktan sonra ancak eşleşen veriler daha fazla analiz için kullanılmalıdır. Eşleme, nedensel bağlantılar hakkında kesinlik sağlamaz ve kendi başına bir müdahalenin etkisi olan mekanizma hakkında iç görü sağlama olasılığı düşüktür (29). Eşleme, özellikle gözlemsel çalışmalarda nedensel çıkarımları iyileştirmek için güçlü, parametrik olmayan bir yaklaşımdır. Uygulamasının basit ve kolay anlaşılabilir olmasından dolayı eşleme, uygulama ile uğraşan araştırmacılar arasında giderek daha popüler hale gelmektedir. Temel fikir, bir veri setindeki bazı ciddi istatistiksel sorunların, çıkarımları dikkatle seçilmiş bir alt veri setiyle sınırlandırarak kaldırılabileceğidir. Eşleme heterojen gözlemleri kaldırarak bazen varyansı azaltabilir, ancak varyans arttığında yanlılığın azaltılması genellikle tipik büyük gözlemsel veri setlerinde telafi edilenden daha fazladır (30). Gözlemsel veriler kullanılarak nedensel etkiler tahmin edilirken, benzer ortak değişken dağılımlarına sahip deneme ve kontrol grupları elde edilerek, randomize bir deneyin mümkün olduğunca benzer bir şekilde tekrarlanması arzu edilmektedir. Bu hedefe genellikle, orijinal deneme ve kontrol gruplarının iyi eşleşmiş örneklerinin seçilmesi ve böylece ortak değişkenlere bağlı yanlılığın azaltılmasıyla ulaşılabilir (31). Mikro veri setlerinin eşleştirilmesi araştırma ve istatistiksel amaçlar için çok yararlıdır. Eşlemenin kullanılmasıyla, alternatif yöntemlere göre daha düşük maliyetle veya daha kısa sürede analizler yapmak veya tahminler yapmak mümkündür. Bazı durumlarda eşleme, araştırma yapmanın tek uygun yoludur. Eşleme yoluyla elde edilen analizler veya tahminler bazen diğer şekillerde elde edilenlerden daha

güvenilirdir. Ayrıca, eşleme çoğu zaman yanıt yükünde bir azalmaya yol açar (8). İstatistiksel eşleme, birden fazla veri seti aracılığıyla toplanan değişkenler ve göstergeler hakkında ortak bilgi kanıtlamaya yönelik model tabanlı bir yaklaşımdır. Bu yaklaşımın potansiyel faydaları, mevcut veri kaynaklarının tamamlayıcı kullanımını ve analitik potansiyelini artırma ihtimalindedir. Dolayısıyla, istatistiksel eşleme, mevcut veri toplamaları göz önüne alındığında kullanım verimliliğini artırmak için bir araç olabilir (7). İstatistiksel eşleme için bazı modeller ve spesifik stratejiler literatürde tartışılmıştır. Tam eşleme veya istatistiksel eşleme stratejileri arasında ayrım yapmak gelenekseldir. Bununla birlikte, dosya eşleme ile ilgili literatürde, tam eşleme ve istatistiksel eşleme tanımları üzerinde bir fikir birliği yoktur (6).

SONUÇ

Rubin tarafından önerilen yöntem, diğer yaklaşımlara göre oldukça iyi sonuçlar vermesine rağmen en iyi yöntem konusunda fikir birliği bulunmamaktadır. Veri kaynakları üzerindeki sınırlar ve eşleme amacı bir yöntem seçiminde çok önemlidir ve bu faktörler eşlemeden eşlemeye farklılık göstereceği için fikir birliği eksikliği şaşırtıcı değildir. Dolayısıyla gerek kısıtlı gerekse kısıtsız eşleme yöntemleri hala kullanılmaktadır.

Yazarların Katkıları: Fikir/Kavram: İ.D., N.D.; Tasarım: İ.D.; Veri Toplama ve/veya İşleme: İ.D., N.D.; Analiz ve/veya Yorum: İ.D., N.D.; Literatür Taraması: İ.D., N.D.; Makale Yazımı: İ.D.; Eleştirel İnceleme: İ.D., N.D.

KAYNAKLAR

1. Marella D, Pfeffermann D. Matching information from two independent informative samples. J Stat Plan Infer. 2019; 203: 70-81.
2. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat. 1986; 4(1): 87-94.
3. Barry RA, Stewart WH, Turner JS. An empirical evaluation of statistical matching methodologies. Dallas, Texas: Working Paper, Southern Methodist University; 1982. p. 4-5.
4. Barry, JT. An investigation of statistical matching. J Appl Stat. 1988; 15(3): 275-83.
5. Willenborg L, Heerschap H. Statistics methods: matching. Hague: Statistics Netherlands; 2012.
6. Goel PK, Ramalingam T. The matching methodology: some statistical properties. Berlin: Springer-Verlag; 1989.
7. European Union (eurostat). Statistical matching: a model based approach for data integration. Luxembourg: Publications Office of the European Union; 2013.
8. Radner DB, Allen R, Gonzalez ME, Jabine TB, Muller HJ. Report on exact and statistical matching techniques. Washington: Government Printing Office; 1980.
9. Ahfock D, Pyne S, Lee SX, McLachlan GJ. Partial identification in the statistical matching problem. Comput Stat Data Anal. 2016; 104: 79-90.
10. D'Orazio M, Zio MD, Scanu M. Statistical matching: theory and practice. West Sussex: John Wiley & Sons Ltd.; 2006.
11. National Statistics. National statistics code of practice protocol on data matching. London: A National Statistics Publication; 2004.
12. Moriarity CL, Scheuren F. Statistical matching: pitfalls of current procedures. Proceedings of the Annual Meeting of the American Statistical Association; 2001; Atlanta, Georgia.
13. Rasner A, Frick JR, Grabka MM. Statistical matching of administrative and survey data: an application to wealth inequality analysis. Sociol Method Res. 2013; 42(2): 192-224.
14. Rodgers WL. An evaluation of statistical matching. J Bus Econ Stat. 1984; 2(1): 91-102.
15. Endres E, Augustin T. Statistical matching of discrete data by Bayesian networks. In: Antonucci A, Corani G, de Campos CP, editors. Pgm2016: Proceedings of the 8th International Conference on Probabilistic Graphical Models; 2016 Sep 6-9; Switzerland. Lugano: 2016. p. 159-70.
16. Rassler S. Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. New York: Springer Science+Business Media, LLC; 2002.
17. Kadane JB. Some statistical problems in merging data files. J Off Stat. 2001; 17(3): 423-33.
18. Wiest M, Kutscher T, Willeke J, Merkel J, Hoffmann M, Kaufmann-Kuchta K, et al. The potential of statistical matching for the analysis of wider benefits of learning in later life. European Journal for Research on the Education and Learning of Adults. 2019; 10(3): 291-306.
19. Conti PL, Marella D, Neri A. Statistical matching and uncertainty analysis in combining household income and expenditure data. Stat Methods Appl. 2017; 26: 485–505.
20. Gessendorfer J, Beste J, Drechsler J, Sakshaug JW. Statistical matching as a supplement to record linkage: a valuable method to tackle nonconsent bias?. J Off Stat. 2018; 34(4): 909-33.
21. Conway A, Rolley JX, Fulbrook P, Page K, Thompson DR. Improving statistical analysis of matched case–control studies. Res Nurs Health. 2013; 36(3): 320-4.
22. Faresjö T, Faresjö A. To match or not to match in epidemiological studies-same outcome but less power. Int J Environ Res Public Health. 2010; 7(1): 325-32.
23. Kim K, Park M. Statistical micro matching using a multinomial logistic regression model for categorical data. CSAM. 2019; 26(5): 507-17.
24. Moriarity CL, Scheuren F. Regression based statistical matching: Recent developments. Proceedings of the Section on Survey Research Method, American Statistical Association, 2004; p. 4050-7.
25. Waal Ton de. Statistical matching: Experimental results and future research questions. Den Haag: CBS. 2015. doi: 10.13140/RG.2.1.1969.4161.
26. Perchinunno P, Mongelli L, d'Ovidio F. Statistical matching techniques in order to plan interventions on

socioeconomic weakness: an Italian case. Socio Econ Plan Sci. 2020; 71. 100836. 10.1016/j.seps.2020.100836.
27. Alpman A, Gardes F, Thiombiano N. Statistical Matching for Combining Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Data. Documents de travail du Centre d'Economie de la Sorbonne 17024, Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne. 2017. ffhalshs-01529699f
28. Namazi-Rad MR, Tanton R, Steel D, Mokhtarian P, Das S. An unconstrained statistical matching algorithm for combing individual and household level geo-specific census and survey data, National Institute for Applied Statistics Research Australia, University of Wollongong, Working Paper 01-16, 2016, 35.
29. 29. Schleicher J, Eklund JD, Barnes M, Geldmann J, Oldekop JA, Jones JPG. Statistical matching for conservation science. Conserv Biol, 2019; doi: 10.1111/cobi.13448.
30. 30. Iacus SM, King G, Porro G. A theory of statistical inference for matching methods in causal research. Political Analysis. 2019; 27(1): 46-68.
31. 31.Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010; 25(1): 1-21.