گزیده‌ای از تازه‌های پزشکی، بیماری‌های داخلی و کودکان
سال دوم، شماره اول، تابستان ۱۴۰۴، صفحات ۵۴ تا ۶۲

اصیل

مدیریت داده‌های گم‌شده در پژوهش‌های بالینی: مفاهیم، چالش‌ها و روش‌های برخورد با آن‌ها با استفاده از نرم‌افزار R

الهام مدرسه*[۱،۲]، نسرین حسینقلی‌زاده[۳]، معصومه اخلاقی[۱،۴]، مجید علیخانی[۱،۴]، شکوفه صادقی[۱،۴]

۱. مرکز تحقیقات روماتولوژی، دانشگاه علوم پزشکی تهران، تهران، ایران
۲. واحد توسعه تحقیقات بالینی، بیمارستان شریعتی، دانشگاه علوم پزشکی تهران، تهران، ایران
۳. دانشکده علوم پزشکی میاندوآب، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران
۴. گروه بیماری‌های داخلی، دانشکده پزشکی، بیمارستان شریعتی، دانشگاه علوم پزشکی تهران، تهران، ایران
*نویسنده مسئول: emadreseh@yahoo.com

چکیده

وجود داده‌های گم‌شده، یکی از چالش‌های شایع و اغلب اجتناب‌ناپذیر در علم داده و پژوهش‌های بالینی محسوب می‌شود. این مسئله می‌تواند بر دقت، اعتبار درونی و تفسیر نتایج پژوهش تأثیرگذار باشد. در این مسیر، درک دقیق مجموعه داده‌ها به تحلیل‌گران سلامت امکان می‌دهد تا راهبردهایی را برای پیشگیری و کاهش داده‌های گم‌شده در مراحل طراحی و اجرای مطالعه به کار گیرند. با این حال، به دلیل ماهیت تحقیقات بالینی، وجود داده‌های ناقص همچنان اجتناب‌ناپذیر است که استفاده از راهکارهای عملی برای مدیریت داده‌های گم‌شده را ضروری می‌سازد. این مقاله، روش‌های اصلی نحوه برخورد با داده‌های گم‌شده را مرور می‌کند و به معرفی انواع سازوکار و الگوی گم‌شدگی و همچنین نسبت داده‌های گم‌شده قابل چشم‌پوشی می‌پردازد. در پایانبا ارائه یک مثال بر روی مجموعه داده فرضی مربوط به بیماری آرتریت روماتویید، یکی از کاربردی‌ترین روش‌های جایگذاری داده‌های گم‌شده (جانهی چندگانه با معادلات زنجیره‌ایی) را معرفی نموده وبا استفاده از بسته mice از نرم‌افزار R، کد های مربوطه اجراو تفسیر می‌گردند. محققان با هر سطح دانش نسبت به علم آمار زیستی و نرم‌افزار R، می‌توانند با اجرای کدهای ضمیمه شده در مقاله حاضر، درصورت برقرار بودن پیش‌فرض‌های مربوطه، به برآورد داده‌های گم‌شده در مجموعه داده‌های پژوهش خود بپردازند.

کلیدواژه‌ها: جانهی آماری، داده‌های گم‌شده، گم‌شدگی غیرتصادفی، یادگیری عمیق، یادگیری ماشین

مقدمه

فقدان برخی اطلاعات در مجموعه داده‌ها به‌ویژه در حوزه سلامت، یکی از چالش‌های تحلیل داده‌های موجود به‌شمار می‌رود. داده‌های گم‌شده (Missing Data) به مقادیری اشاره دارد که در یکیا چند متغیر ازیک مجموعه داده، مشاهده نمی‌شوند ولی در صورت موجود بودناین مقادیر، نقش مهمی در تحلیل و نتیجه‌گیری خواهند داشت. بنابراین، هر مقدار گم‌شده، نمایانگر بخشی از اطلاعات ارزشمندی است که در دسترس نمی‌باشد(۳-۱).وجود داده‌های ناقص می‌تواند به دلایل مختلفی در زمان جمع‌آوری، ذخیره‌سازی یا پردازش داده‌ها ایجاد شود و یا در برخی موارد، داده‌های ناقص ممکن است شامل نقاط پرت (Outliers) باشند(۴). به طور مثال، شرکت‌کنندگان در یک نظرسنجی ممکن است، تمایلی به پاسخ‌گویی به برخی سؤالات نداشته باشند یا اعداد آزمایشگاهی به صورت غیرمنطقی خیلی بزرگ یا خیلی کوچک باشند.از طرف دیگر از آنجا که دقت و کارایی تحلیل‌های آماری به‌طور قابل توجهی به کیفیت داده‌ها وابسته است، وجود داده‌های گم‌شده، تحلیل داده‌ها را پیچیده کرده و در صورت حذف این موارد، حجم نمونه کاهش می‌یابد که می‌تواند منجر به کاهش توان آماری و ایجاد سوگیری در تخمین اثرات درمان یا مداخله و کاهش دقت فواصل اطمینان شود(۳، ۵، ۶). همچنین این مسئله، بر دقت پیش‌بینی‌هااثرگذار بوده و فرآیندهای تصمیم‌گیریبرای سیاستگذاران را دچار سوگیریمی‌نماید(۶، ۷).

ساختار مقادیرگم‌شده به سه مفهوم، نرخ داده‌های گم‌شده (Missing Data Rate)، الگوی داده‌های گم‌شده (Missing Data Pattern) و سازوکار(مکانیسم) داده‌های گم‌شده (Missing Data Mechanism) در یک مجموعه داده اشاره دارد.نرخ گم‌شدگی (نسبت داده‌های ازدست رفته به کل مجموعه داده)، نقش مهمی در نحوه تحلیل داده‌ها ایفا می‌کند. هرچه نرخ گم‌شدگی بالاتر باشد (بیش از ۱۵ درصد)، تحلیل داده‌ها را دشوار و

گاهی غیرممکن می‌سازد(۸). الگوی داده‌های گم‌شده به نحوه توزیع داده‌های گم‌شده از نظر ظاهری، در یک مجموعه داده اشاره دارد و مشخص می‌کند که کدام متغیرها در کدام مشاهدات فاقد مقدار هستند. در حالی که سازوکار گم‌شدگی به فرآیند و دلایلی اشاره دارد که باعث ایجاد مقادیر از دست رفته در یک مجموعه داده می‌شود(۱). در ادامه دو مفهوم اخیر مفصل‌تر توضیح داده می‌شوند.

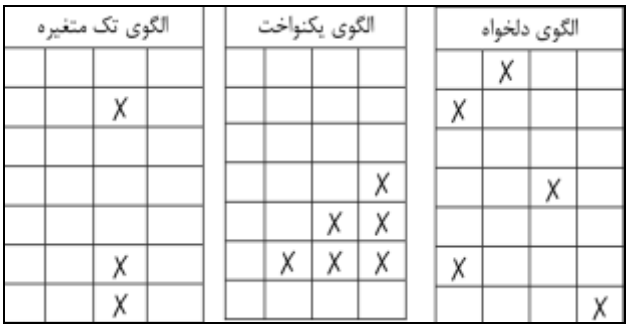
انواع الگوی گم‌شدگی

به طور کلی در منابع سه الگوی مختلف گم‌شدگی نام برده شده است:

- **الگوی گم‌شدگی تک‌متغیره (Univariate Pattern):** فقط یک متغیر یا یک ستون داده‌ها، مقادیرگم‌شده دارند؛ مثلاً سن شروع بیماری برای برخی افراد در دسترس نیست، اما مقادیر سایر متغیرها در مجموعه داده موجود است
- **الگوی گم‌شدگی یکنواخت (Monotone Pattern):** اگر مقدار یک متغیر خاص نامعلوم باشد، مقادیر متغیرهای بعدی هم نامعلوم می‌شوند؛ اغلب در داده‌های طولی یا در پیگیری بیماران این نوع الگو مشاهده می‌شود؛ مثلاً عدم ثبت اطلاعات فرد از یک معاینه به بعد.
- **الگوی گم‌شدگی عمومی/ دلخواه (Arbitrary Pattern):** داده‌های گم‌شده به صورت پراکنده و بدون نظم خاص در مجموعه داده ظاهر می‌شوند، بیشتر در سامانه‌های ثبت داده‌ها این الگو قابل مشاهده است؛ به عنوان مثال، در برخی بیماران داده‌های برخی آزمایش‌ها موجود نیست و الگوی مشخصی در گم‌شدگی نیز دیده نمی‌شود (شکل۱)(۴).

انواع سازوکار گم‌شدگی

در متون علمی سه سازوکار برای داده‌های گمشده، تعریف می‌گردد: الف) گم‌شدگی کاملاً تصادفی (Missing completely at Random; MCAR)، ب) گم‌شدگی تصادفی (Missing at Random; MAR) و ج) گم‌شدگی غیرتصادفی (Missing not at Random; MNAR) یا آگاهی‌بخش (Informative Missing Values) که هر یک چالش‌های منحصربه‌فردی را در فرآیند جانهی (پر کردن مقادیر گم‌شده) ایجاد می‌کنند(۱).



شکل ۱. الگوهای داده‌های گم شده در مجموعه داده‌ها؛ علامت X نشان‌دهنده داده‌های گم‌شده

- **گم‌شدگی کاملاً تصادفی (MCAR):** در این حالت، مفقود شدن داده هیچ ارتباطی با سایر متغیرهای مشاهده‌شده یا مشاهده‌نشده ندارد. به عبارت دیگر، بین افرادی که داده‌هایشان مفقود است و آن‌هایی که داده‌های کامل دارند، تفاوت نظام‌مندی وجود ندارد. برای مثال، ممکن است مقادیر آزمایشگاهی برخی شرکت‌کنندگان به دلیل خطا در پردازش دستگاه مفقود شده باشد و یا در یک سامانه ثبت داده‌ها که میزان افسردگی افراد را ثبت می‌کند، کاربرفراموش کرده باشد که پرسشنامه افسردگی را در اختیار بیمار قرار دهد ومیزان افسردگی ثبت نگردد. در چنین شرایطی، اگر نرخ گم‌شدگی کمتر از ۱۵-۱۰ درصد باشد، با حذف افرادی که داده‌های افسردگی برای آن‌ها ثبت نشده و انجام تحلیل‌ها بر روی سایر افراد یعنی موارد کامل (Complete cases)، اگرچه حجم نمونه کاهش یافته و در نتیجه توان آماری نیز کم می‌شود، اما سوگیری ایجاد نمی‌شود (دقت برآوردها کاهش می‌یابد اما همچنان قابل اعتماد است)؛ زیرا داده‌های باقی‌مانده را می‌توان به‌عنوان نمونه‌ای تصادفی از کل داده‌ها در نظر گرفت(۴).
- **گم‌شدگی تصادفی (MAR):** در این حالت، مفقود شدن داده‌ها به‌طور نظام‌مند با داده‌های مشاهده‌شده مرتبط است، اما با داده‌های مشاهده‌نشده ارتباطی ندارد. در مثال قبل، با فرض گم‌شدگیتصادفی، ممکن است مردان کمتر از زنان پرسشنامه شدّت افسردگی را تکمیل کنند. به این معنا که احتمال تکمیل پرسشنامه به جنسیّت (که داده‌ای کاملاً مشاهده‌شده است) بستگی دارد، اما به شدّت افسردگی (داده مشاهده‌نشده) ربطی ندارد. در این شرایط تحلیل موارد کامل (یعنی داده‌هایی که تمام اطلاعات مربوطه را دارند) ممکن است سوگیری داشته یا نداشته باشد. در صورت وجود سوگیری، مدل‌های آماری مناسب می‌توانند با لحاظ کردن عوامل شناخته‌شده (مانند جنسیّت در این مثال)، تحلیلی بدون سوگیری انجام دهند(۴).
- **گم‌شدگی غیرتصادفی (MNAR):**در این حالت، مفقود شدن داده‌ها به‌طور نظام‌مند به داده‌های مشاهده‌نشده مربوط است، یعنی مفقود شدن داده‌ها به عوامل یا رویدادهایی وابسته است که توسط پژوهشگر اندازه‌گیری نشده‌اند و این عوامل با پیامد اصلی نیز ممکن است مرتبط باشند. در مثال مذکور، اگر افراد با افسردگی شدیدتر، از تکمیل پرسشنامه خودداری کنند، ممکن است الگوی گم‌شدگی غیرتصادفی باشد. در این حالت مانند الگوی گم‌شدگی تصادفی تحلیل موارد کامل ممکن است سوگیری داشته یا نداشته باشد. اگر سوگیری وجود داشته باشد، از آنجا که عوامل ایجادکننده مفقود شدن داده‌ها اندازه‌گیری نشده‌اند، لازم است از مدل‌های

پیشرفته آماری (مانند مدل‌های توام (Joint Models)) برای یک تحلیل بدون سوگیری استفاده کرد(۴).

در زمان تحلیل داده‌ها و انتخاب نوع روش آماری، بررسی سازوکار گم‌شدگی از اهمیت بسیاری جهت دستیابی به نتایج قابل اعتماد دارد. این مسئله به ویژه در مورد متغیرهای مستقل پژوهش مورد توجه قرار می‌گیرد. لازم به ذکر است که هیچ کدام از این سه سازوکار از نظر آماری قابل آزمون نمی‌باشد و تنها با تحلیل حساسیت قابل تشخیص می باشند. بدین صورت که ابتدا با حذف موارد ناقص، داده‌هاتحلیل می‌شوند. سپس با روش‌های جانهی متناسب با هر سازوکار، داده‌های گم‌شده برآورد شده و همراه سایر داده‌هاتحلیل می‌شوند. سرانجام در صورتی که نتایج هر دو رویکرد (با و بدون برآورد مقادیر گمشده) یکسان باشد، بیانگر سازوکار گم‌شدگی کاملا تصادفی است،بنابراین در این حالت حذف موارد گم‌شده سوگیری ایجاد نمی‌کند.

روش‌های برخورد با مقادیر گم شده

در این قسمت به برخی روش‌های متداول که قادر به مدیریت انواع سازوکارهای داده‌های گم شده هستند، اشاره می‌شود. علاقه‌مندان برای دریافت جزئیاتمطلب به منابع مربوطه ارجاع داده می‌شوند. به طور کلی روش‌های مدیریت مقادیر گم شده شامل: روش‌های حذف (Deletion)، روش‌های جایگزینی (Imputation)، و یادگیری بازنمایی (Representation Learning) می‌باشند(۹).

روش‌های حذف: ساده‌ترین راهکار برای برخورد با داده‌های ناقص حذف کردن می‌باشد. در این رویکرد، ردیف‌ها یا ستون‌هایی که شامل مقادیر گم شده هستند، از مجموعه داده کنار گذاشته می‌شوند. هرچند این روش‌ها از نظر درک و اجرا آسان هستند، امازمانی که داده‌های گم شده از الگو یا سازوکار خاصی تبعیت می‌کنند، می‌تواند منجر به نتایج جانبدارانه شوند. علاوه بر این، حذف بی‌رویه داده‌ها ممکن است منجر به از دست رفتن اطلاعات ارزشمند و ایجاد سوگیری در تحلیل‌های آماری بعدی شود.

روش‌های جایگزینی: به منظور رفع محدودیت‌های روش‌های حذف، روش‌های جایگزینی نقش مهمی در مدیریت داده‌های ناقص ایفا می‌کنند. هدف اصلی این روش‌ها، بازسازی مقادیر گم شده با حفظ یکپارچگی مجموعه دادهاست. این رویکردها به‌ویژه زمانی سودمند هستند که داده‌های ناموجود محدود باشند یا سازوکار خاصی موجب گم‌شدگی داده‌ها شود. روش‌های جایگزینی شامل تکنیک‌های متنوعی مانند جایگزینی با میانگین/میانه/مد(۹، ۱۰)، اخرین مقدار مشاهده شده (Last Observation Carried Forward; LOCF)، مقدار مشاهده‌شده بعدی (Next Observation Carried Backward; NOCB) (۱۱)،جایگزینی مبتنی بر رگرسیون (جانهی چندگانه، ماکزیمم درستمایی (Maximum Likelihood)، رویکرد بیزی (Bayesian Approach) و ...)(۱۴-۱۲)، رویکردهای مبتنی بر یادگیری ماشین (نزدیک‌ترین همسایه (K-Nearest Neighbors; KNN)، درخت تصمیم (Decision Tree)، ماشین بردار پشتیبان (Support Vector Machine)، خوشه‌بندی (Clustering) و ...)(۱۸-۱۵)، شبکه عصبی (Neural Network) (۱۹) و ... هستند. این روش‌ها با بهره‌گیری از اطلاعات موجود در داده‌های مشاهده‌شده، امکان انجام تحلیل‌های جامع‌تر و کاهش سوگیری ناشی از حذف داده‌ها را فراهم می‌کنند.در محیط نرم‌افزارR ، پکیج miceیکی از ابزارهای جامع و پرکاربرد برای پیاده‌سازی برخی از روش‌های فوق محسوب می‌شود (۱۴).

روش‌های یادگیری بازنمایی: این روش‌یکی از مفاهیم کلیدی در حوزه هوش مصنوعی و یادگیری عمیق می‌باشد که به کمک آن می‌توان اطلاعات پنهان داده‌ها را استخراج و به جای اتکا صرف بر داده‌های هر متغیر به طور جداگانه، از ویژگی‌ها و ساختارهای یادگرفته‌شده برای ارتقای کیفیت و دقت مقادیر جایگزین بهره برد. در ادامهبه دو روش برای مدیریت داده‌های گم شده در این حوزه اشاره می‌گردد.

شبکه‌های عصبی گراف (Graph Neural Networks; GNNs): این روش با درنظر گرفتن مجموعه داده‌ها به صورت یک گراف، قادر به بازسازی مقادیر گم شده می‌باشد. در این گراف هر نمونه یا بیمار به عنوان یک گره (Node)، روابط بین آن‌ها به عنوان یال (Edge) مدل‌سازی می‌شوند. در ادامه شبکه با یافتن الگوی وابستگی‌های بین اطلاعاتافرادی که با هم در گراف متصل شده‌اند، مقادیر گمشده را با دقت مناسبی تخمین می‌زند.

خودرمزگذارها (Autoencoders): این روش نوعی شبکه عصبیمی‌باشد که مجموعه داده ورودیکه ممکن است شامل داده‌های گمشده نیز باشند را فشرده کرده و به یک فضای با ابعاد کمترتبدیل می‌کند، سپس شبکه با یادگیری و کشف ویژگی‌های پنهان، امکان بازسازی داده‌های ورودی و در نتیجه برآورد مقادیر ناقص را فراهم می‌آورند. پیاده‌سازی روش اخیر در محیط نرم‌افزار R از طریق پکیج rMIDAS امکان‌پذیر است(۲۲-۲۰).

نتیجه‌گیری

داده‌های ناقص یکی از چالش‌های رایج در تحلیل داده‌های بالینی و پژوهش‌های اپیدمیولوژیک می‌باشند. همان‌طور که اشاره شد، سازوکارهای گم‌شدگی می‌توانند به صورت MCAR، MAR و MNAR باشند که هر یک پیامدهای متفاوتی بر تحلیل و سوگیری نتایج دارند. همچنین، نرخ گم‌شدگی نقش تعیین‌کننده‌ای در انتخاب استراتژی مناسب ایفا می‌کند. در مواردی که درصد گم‌شدگی کم باشد، حذف نمونه‌ها، ممکن است کافی باشد، اما با افزایش نرخ گم‌شدگی، استفاده از روش‌های جایگزینی یا مدل‌سازی‌های پیچیده ضروری می‌باشد.در مجموع، انتخاب روش مناسب برای مدیریت داده‌های گم شده باید بر اساس سازوکارگم‌شدگی، نرخ داده‌های گم شده و نوع داده‌ها صورت گیرد. رویکردهای پیشرفته مبتنی بر یادگیری ماشین و یادگیری نمایش مسیر نویدبخشی برای مقابله با

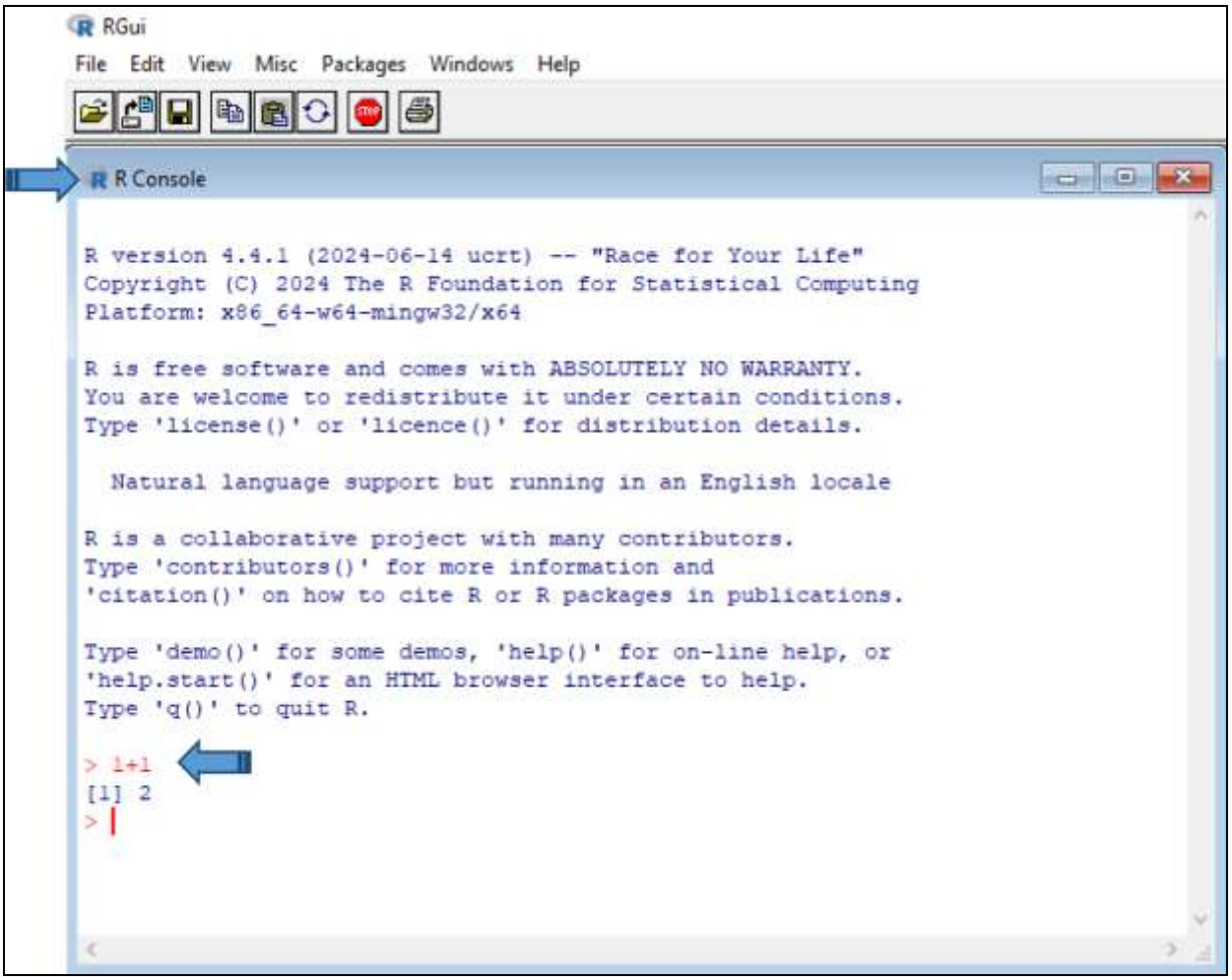
چالش‌های داده‌های ناقص ارائه می‌دهند و می‌توانند به تحلیل‌های بالینی و پژوهشی دقیق‌تر و قابل اعتمادتر منجر شوند.

پیوست: مثال کاربردی در نرم‌افزار R

این مثال از دو بخش اصلی تشکیل شده است. در بخش اول، با نوشتن کدهای مربوطه در محیط نرم‌افزار R، یک مجموعه داده فرضی ایجاد و در بخش دوم، روش‌های جایگذاری داده‌های گمشده بر روی این داده‌ها نشان داده خواهد شد.

ابتدا از طریق تارنما https://cran.r-project.org/bin/windows/base/ آخرین نسخه از نرم‌افزار R دانلود و سپس بر روی رایانه قایل نصب می‌باشد.

پس از نصب، با دبل کلیک روی آیکن نرم افزار، آماده اجرای دستورات و کدهای تایپ شده در R console می‌شود.



گام اول: تولید و شبیه‌سازی داده‌های فرضی برای یک مطالعه کارآزمایی بالینی در بیماران آرتریت روماتویید

فرض کنید قصد داریم داده‌های مربوط به یک کارآزمایی بالینی تصادفی‌شده (RCT) را روی بیماران مبتلا به آرتریت روماتویید (RA) شبیه‌سازی کنیم. هدف این مطالعه بررسی اثر یک درمان جدید بر میزان ناتوانی بیماران است. برای این منظور، از بسته lava در نرم‌افزار R استفاده می‌کنیم. متغیر وابسته در این مدل، نمره پرسشنامه ارزیابی سلامت (HAQ) است که دامنه‌ای بین صفر (بدون ناتوانی) تا ۳ (ناتوانی شدید) دارد. سایر متغیرهای مورد استفاده شامل نوع درمان، مدت زمان

بیماری و نرخ رسوب گلبول قرمز (ESR) هستند.پس از نصب پکیج lava و فرخوانی آن و با اجرای کدهای زیر در نرم‌افزار R، یک مجموعه داده با حجم نمونه ۲۰۰ بیمار و به نام data.RA ایجاد می‌شود. set.seed(123) باعث می‌شود در هر بار تولید داده‌ها یکسان باشد. در پایان با دستور head(data.RA)، شش سطر اول داده‌هامشاهده می‌شود:

```
> install.packages("lava")
    :
> library(lava)
> mSim.RA <- lvm(haq ~ treatment + duration + esr + gender + age)
> categorical(mSim.RA, labels = c("Placebo", "Biologic")) <- ~treatment
> categorical(mSim.RA, labels = c("Male", "Female")) <- ~gender
> distribution(mSim.RA, ~age) <- lava::uniform.lvm(40, 85)
> distribution(mSim.RA, ~duration) <- lava::gaussian.lvm(mean = 8, sd = 5)
> distribution(mSim.RA, ~esr) <- lava::gaussian.lvm(mean = 25, sd = 15)
> set.seed(123)
> data.RA <- sim(mSim.RA, n = 200)
> head(data.RA)
       haq treatment   duration      esr gender      age
1  91.33916  Biologic 13.3700613 30.34425   Male 47.18533
2  70.26615  Biologic  7.8632652 15.12985   Male 46.50321
3  93.93321   Placebo  7.8333483 37.82803   Male 46.71312
4 106.93376  Biologic  0.4196619 42.29404   Male 63.14954
5 104.40256   Placebo 11.9519267 29.14412 Female 62.17723
6 103.55839   Placebo  6.9463291 27.16157   Male 67.73542
>
```

در ادامه، به ترتیب پنج، ده و پانزده مورد از مقادیر متغیرهای HAQ، مدت زمان بیماری(duration) و ESR را به طور تصادفی حذف خواهیم کرد (منظور از NA در نرم‌افزار R، داده گمشده است). در خط پایانی کدهای زیر، نرخ گمشدگی هر متغیر محاسبه و نمایش داده شد. از آنجا که میزان گمشدگی همه متغیرها کمتر از ۱۵ درصد است و با فرض اینکه سازوکار گمشدگی کاملاً تصادفی است، از روش جانهی چندگانه، داده‌های گمشده برآورد خواهند شد. در مرحله بعد، از این مجموعه داده جدید که شامل داده گمشده است، استفاده خواهیم کرد و نشان می‌دهیم که چگونه می‌توان با استفاده از مدل رگرسیونی و سایر متغیرها، این مقادیر گمشده را برآورد کرد:

```
> data.RA$esr[sample(1:nrow(data.RA), 5)] <- NA
> data.RA$haq[sample(1:nrow(data.RA), 10)] <- NA
> data.RA$duration[sample(1:nrow(data.RA), 15)] <- NA
> # محاسبه و نمایش درصد گمشدگی برای هر متغیر
> colMeans(is.na(data.RA)) * 100
     haq treatment  duration       esr    gender       age
     5.0       0.0       7.5       2.5       0.0       0.0
>
```

گام دوم: جایگذاری مقادیر گمشده با استفاده از بسته نرم‌افزاری mice

پس از ایجاد داده‌های فرضی (یا درصورت وجود مجموعه داده واقعی)، اکنون نوبت به مدیریت داده‌های گمشده می‌رسد. با تایپ دستور fix(data.RA) کل مجموعه داده‌ها به صورت زیر قابل مشاهده است. در مجموعه داده حاضر، مقادیر مربوط به نمره ناتوانی (HAQ)، مدت زمان بیماری (duration) و نرخ رسوب گلبول قرمز (ESR) برای برخی از بیماران ثبت نشده است و به صورت NA نمایش داده می‌شود.

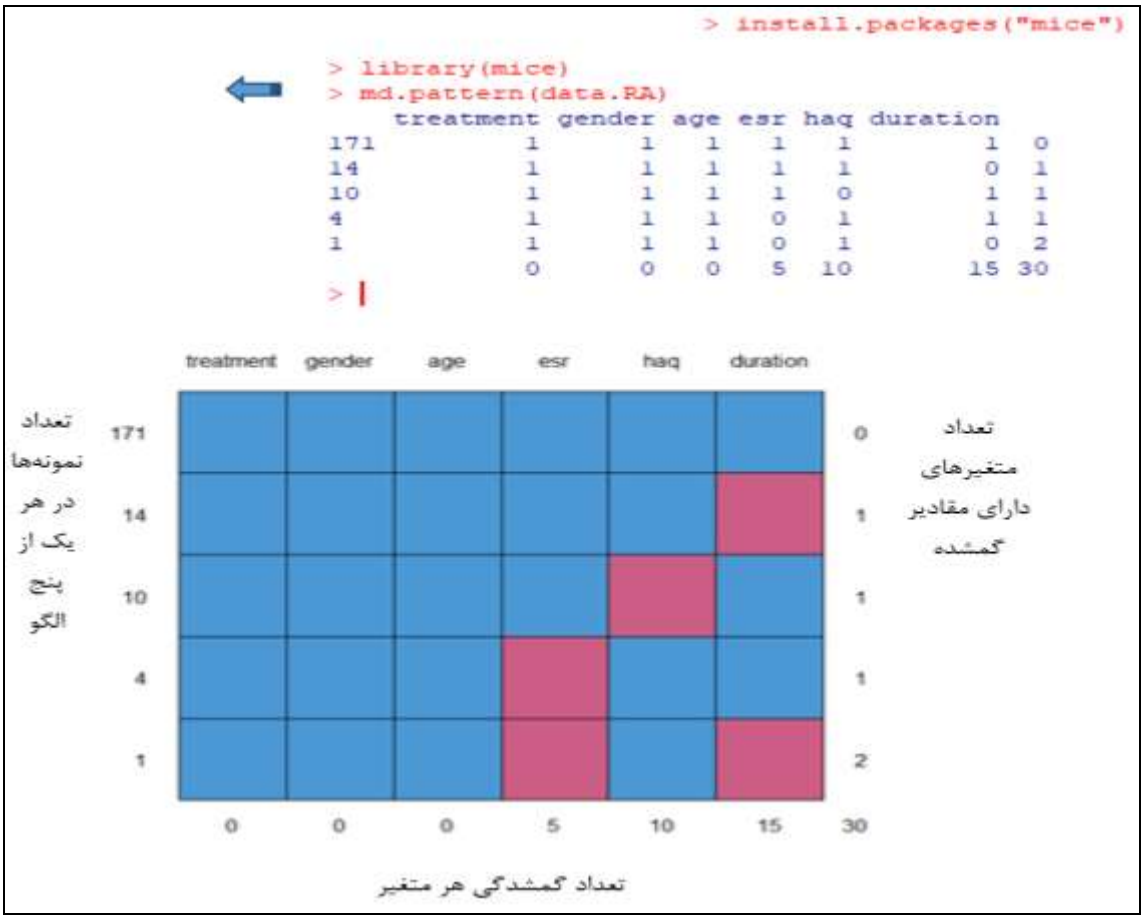
در ادامه با دستور install.packages("mice") بسته mice نصب و با دستور library(mice) این بسته فراخوانی می‌شود. با استفاده از دستور md.pattern(data.RA) الگوی گمشدگی مجموعه داده‌ها در شکل۲ مشاهده می‌گردد. بر این اساس پنج الگو برای افراد تحت مطالعه مشاهده می‌شود. الف) ۱۷۱ نفر، مقادیر همه متغیرهایشان ثبت شده و بدون مقادیر گمشدگی هستند (موارد کامل). ب) ۱۴ نفر فقط طول مدت بیماریشان ثبت نشده است. پ) ۱۰ نفر فقط haq برایشان ثبت نشده است. ت) چهار نفر فقط ESR برایشان ثبت نشده و نهایتا ث) یک نفر ESR و duration ندارد.

```
> fix(data.RA)
```

Data Editor

	haq	treatment	duration	esr	gender	age	var7
51	99.56251	Placebo	NA	33.43401	Female	66.5677	
52	77.39413	Placebo	8.535952	23.53881	Female	44.34791	
53	135.0379	Placebo	11.0439	40.24683	Female	82.79006	
54	59.09541	Biologic	NA	7.657489	Female	47.32344	
55	153.458	Biologic	10.40313	59.8129	Female	81.46778	
56	NA	Placebo	3.859129	15.94703	Male	45.74988	
57	86.49305	Biologic	13.10127	3.117259	Female	69.82328	
58	107.2951	Biologic	10.69241	19.73623	Male	75.28189	
59	114.5773	Placebo	11.84526	27.20063	Female	74.40759	
60	128.4395	Placebo	8.603597	49.35432	Male	70.26566	
61	97.57099	Placebo	12.31824	38.66815	Female	45.20497	
62	119.8457	Placebo	14.90257	27.13688	Male	78.30857	
63	91.46405	Biologic	17.83124	4.157747	Female	67.80827	
64	97.4874	Biologic	7.858025	12.00943	Male	77.63852	
65	66.88045	Biologic	-3.245255	22.55073	Male	47.64677	
66	114.2192	Biologic	NA	NA	Female	40.46265	
67	62.12874	Placebo	9.027806	-2.903414	Female	54.55614	
68	102.3778	Placebo	7.223273	41.96582	Female	52.13572	
69	106.992	Placebo	10.84144	17.09149	Female	77.13676	

```
> install.packages("mice")
> library(mice)
> md.pattern(data.RA)
    treatment gender age esr haq duration
171         1      1   1   1   1        1 0
14          1      1   1   1   1        0 1
10          1      1   1   1   0        1 1
4           1      1   1   0   1        1 1
1           1      1   1   0   1        0 2
            0      0   0   5  10       15 30
>
```



شکل ۲. الگوی گمشدگی مجموعه داده فرضی RA

در ادامه با استفاده از متغیرهای مدنظر (مانند سن، جنسیت، طول مدت بیماری، haq و نوع درمان)، مقادیر گمشده از طریق تابع (mice) پیش‌بینی و جایگذاری می‌گردد. خروجی این تابع را به دلخواه imp نامگذاری می‌کنیم.

در تابع (mice)،گزینه m = ۵ باعث می‌شود که پنج دیتاست مجزا از روی data.RA ساخته شود که در هر کدام مقادیر گمشده با اعداد متفاوتی برآوردمی‌شوند. این کار باعث می‌شود در زمان انجام تحلیل‌های آماری، نتایج به دست آمده از تمام این مجموعه داده‌ها (دیتاست‌ها) با هم ترکیب شوند. بنابراین برآوردهای نهایی نسبت به زمانی که فقط از یک بار جایگذاری (m = ۱) استفاده می‌شود، دقیق‌تر می‌شوند.در این مثال ازآنجا که تمام متغیرهای مورد بررسی، کمّی- پیوسته هستند، از گزینه method = 'pmm' جهت برآوردیابی استفاده گردید. برای اطلاع از سایر روش‌ها و همچنین روش‌های جانهی در مطالعات طولی، از دستور help(mice) در محیط نرم‌افزار R می‌توان استفاده نمود.

با استفاده از دستور summary(imp)، تعداد مجموعه داده‌های جانهی شده (m)، روش جانهی (pmm) و ماتریس پیش‌بینی‌کننده (PredictorMatrix)گزارشداده می‌شود. ماتریس پیش‌بینی‌کننده نشان می‌دهد که برای برآورد داده‌های گمشده مربوط به هر یک از متغیرهایی که در سطر ماتریس هستند، از کدام متغیرهای دیگر (در ستون ماتریس) استفاده می‌شود. متغیرهای مورد استفاده با کد ۱ مشخص شده‌اند. در این مثال، جهت پیش‌بینی haq از متغیرهای treatment، duration، ESR، gender و age استفاده می‌شود. این ماتریس قبل از اجرای تابع (mice) طبق نظر محقق قابل تعریف می‌باشد و سپس از طریق گزینه PredictorMatrix در تابع mice به نرم‌افزار معرفی می‌گردد.

با اجرای کد زیر فرآیندهای فوق اجرا می‌گردد:

```
> imp <- mice(data.RA, m = 5, method = 'pmm', seed = 123,maxit = 10,printFlag = FALSE)
> summary(imp)
Class: mids
Number of multiple imputations:  5
Imputation methods:
      haq treatment  duration       esr    gender       age
    "pmm"        ""     "pmm"     "pmm"        ""        ""
PredictorMatrix:
           haq treatment duration esr gender age
haq          0         1        1   1      1   1
treatment    1         0        1   1      1   1
duration     1         1        0   1      1   1
esr          1         1        1   0      1   1
gender       1         1        1   1      0   1
age          1         1        1   1      1   0
>
```

برای مشاهده داده‌های جانهی شده مربوط به هر یک از متغیرها، در هر یک از پنج مجموعه داده از دستور imp$imp استفاده می‌گردد.

در پایان، تحلیل یا مدل آماری مدنظر که هدف پژوهش را پاسخ می‌دهد، بر روی تمام مجموعه داده‌های (دیتاست‌ها) تولید شده اجرا (با تابع (with) و نتایج با هم ادغام می‌گرددبا تابع (pool). برای مثال فرض کنید هدف مطالعه حاضر، تعیین تأثیر مداخله درمان بیولوژیک جدید (treatment) بر میزان ناتوانی (haq) بیماران مبتلا به RA پس از تعدیل اثر مخدوش‌گرهای بالقوه (سن، جنس، طول مدت بیماری و ESR) باشد. با اجرای کد زیر در محیط R، هدف مذکور پاسخ داده می‌شود (تابع (lm) مدل رگرسیون خطی را اجرا می‌کند):

```
> fit <- with(imp, lm(haq ~ treatment + duration + esr + gender + age))
> pooled_fit <- pool(fit)
> summary(pooled_fit)
              term   estimate   std.error  statistic        df      p.value
1      (Intercept) -0.2464059 0.509542038  -0.483583  24.41449 6.329908e-01
2 treatmentBiologic  1.0284270 0.168350624   6.108840  81.59996 3.255254e-08
3         duration  1.0006467 0.017252293  58.000794  57.88984 5.512407e-53
4              esr  0.9964814 0.006565250 151.781176  19.10592 6.563948e-31
5     genderFemale  0.9751093 0.157380155   6.195885 176.15586 3.982546e-09
6              age  1.0059215 0.007402368 135.891865  21.25469 9.885050e-33
>
```

منابع

1. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2019.
2. Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. (No Title). 1985.
3. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
4. Zhou Y, Aryal S, Bouadjenek MR. A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms. arXiv preprint arXiv:240404905. 2024.
5. Demirtas H. Flexible imputation of missing data. Journal of statistical software. 2018;85:1-5.
6. Graham JW, Cumsille PE, Elek-Fisk E. Methods for handling missing data. Handbook of psychology. 2003:87-114.
7. Fan J, Han F, Liu H. Challenges of Big Data Analysis. *Natl Sci Rev*. 2014 Jun;1(2):293-314.
8. Junaid KP, Kiran T, Gupta M, Kishore K, Siwatch S. How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. *Popul Health Metr*. 2025;23(1):2.
9. Little RJ, Rubin DB. Bayes and multiple imputation. Statistical analysis with missing data. 2002:200-220.
10. McGuire S. IOM (Institute of Medicine) and NRC (National Research Council). 2013. Supplemental nutrition assistance program: examining the evidence to define benefit adequacy. Washington, DC: The National Academies Press, 2013. *Advances in Nutrition*. 2013;4(4):477-478.
11. Lachin JM. Fallacies of last observation carried forward analyses. *Clinical trials*. 2016;13(2):161-168.
12. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological). 1977;39(1):1-22.
13. Takahashi M. Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*. 2017;16:37-.
14. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*. 2011;45:1-67.
15. Liao SG, Lin Y, Kang DD, Chandra D, Bon J, Kaminski N, et al. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*. 2014;15(1):346.
16. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. *Stat Anal Data Min*. 2017 Dec;10(6):363-377.
17. Pelckmans K, De Brabanter J, Suykens JA, De Moor B. Handling missing values in support vector machine classifiers. *Neural Networks*. 2005;18(5-6):684-692.
18. Zhang C, Qin Y, Zhu X, Zhang J, Zhang S. Clustering-based missing value imputation for data preprocessing. 2006 4th IEEE International Conference on Industrial Informatics; 2006: IEEE.
19. Chang Y-W, Natali L, Jamialahmadi O, Romeo S, Pereira JB, Volpe G. Neural network training with highly incomplete medical datasets. Machine Learning: Science and Technology. 2022;3(3):035001.
20. You J, Ma X, Ding Y, Kochenderfer MJ, Leskovec J. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*. 2020;33:19075-19087.
21. Bianchi FM, Livi L, Mikalsen KØ, Kampffmeyer M, Jenssen R. Learning representations of multivariate time series with missing data. *Pattern Recognition*. 2019;96:106973.
22. Lall R, Robinson T. Efficient multiple imputation for diverse data in python and r: Midaspy and rmidas. *Journal of Statistical Software*. 2023;107:1-38.

Original

Management of Missing Data in Clinical Research: Concepts, Challenges, and Methods to Address Them Using R Software

Elham Madreseh*[1,2], Nasrin Hosseingholizadeh[3], Maassoumeh Akhlaghi[1,4], Majid Alikhani[1,4], Shokufe Sadeghi[1,4]

1. Rheumatology Research Center, Tehran University of Medical Sciences, Tehran, Iran
2. Clinical Research Development Unit, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran
3. Miandoab Faculty of Medical Sciences, Urmia University of Medical Sciences, Urmia, Iran
4. Department of Internal Medicine, Faculty of Medicine, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran

⁎**Corresponding Author:** emadreseh@yahoo.com

Abstract

The presence of missing data is regarded as one of the most common and frequently unavoidable challenges in data science and clinical research. This issue may adversely affect the accuracy, internal validity, and interpretation of research findings. In this context, an in-depth understanding of datasets enables health data analysts to implement strategies aimed at preventing and minimizing missing data during the design and conduct phases of a study. Nevertheless, owing to the inherent nature of clinical research, incomplete data remain unavoidable, thereby necessitating the use of practical and robust approaches for managing missing data. This article reviews the primary methods for addressing missing data and presents various missing-data mechanisms and patterns, as well as the proportion of missing data that may be considered ignorable. Finally, through an example based on a hypothetical dataset related to rheumatoid arthritis, one of the most widely used approaches for imputing missing data—multiple imputation by chained equations—is introduced. The corresponding codes are implemented and interpreted using the mice package in R software. Researchers with varying levels of expertise in biostatistics and R software can estimate missing data in their research datasets by running the codes attached in this article, provided the relevant assumptions are met.

Keywords: Deep Learning, Machine Learning, Missing Data, Missing not at Random, Statistical Imputation

Please cite this article as follows:
Madreseh E, Hosseingholizadeh N, Akhlaghi M, Alikhani M, Sadeghi Sh. Management of Missing Data in Clinical Research: Concepts, Challenges, and Methods to Address Them Using R Software. Selec Intern Dis and Pediat 2026; 3(1): 54-62.