رسیدگی به داده های از دست رفته در مطالعات اپیدمیولوژی مولکولی

چکیده

مطالعات اپیدمیولوژی مولکولی با مشکل داده های از دست رفته روبروست چرا که نمونه زیستی یا داده های تصویری اغلب تنها در یک نسبت از افراد مورد مطالعه واجدالشرایط برای مطالعه جمع اوری می شود. ما کلیه مطالعات منتشره اپیدمیولوژی مولکولی را مانند مقالات تحقیقاتی، مکاتبات کوتاه، یا نتایج بیطرف خلاصه شده در مجله Cancer Epidemiology, Biomarkers & Prevention را از تاریخ 1 ژانویه 2009 تا 31 مارس 2010 برای مشخصه سازی گستره وجود داده های از دست رفته و روشن سازی روش مطرح سازی این مسئله مورد تحقیق قرار دادیم. از میآن 278 مطالعه اپیدمیولوژی مولکولی ارزیابی شده، بیشتر آنها یعنی 95 درصد دارای داده های از دست رفته درمورد متغیر کلیدی (66٪) بودند یا اینکه از موجودیت داده ها (اغلب اما نه همیشه داده های بیومارکر) به عنوآن معیار ورود به مطالعه (45درصد) استفاده کرده بودند. علی رغم این مورد، تنها ده درصد این مطالعات افراد مورد مطالعه را که در آنالیز گنجانده بودند با انهایی که از آنالیز خارج سازی کرده بودند، مقایسه کرده اند و 88 درصد با داده های از دست رفته یک آنالیز موردی کامل را انجام دادند ، که یک روشی است که برای کسب تخمین های سوگیرانه و ناکارامد هنگامی که داده ها کاملا به طور تصادفی از دست نرفته اند، بکار می رود. یافته های ما شواهدی را فراهم می کند مبتنی براینکه روشهای داده های از دست رفته در مطالعات اپیدمیولوژی مولکولی کمتر از حد مورد استفاده قرار می گیرد که ممکن است اثر بدی روی تفسیر نتایج داشته باشد. ما خط مشی های عملی برای تحلیل وتفسیر مطالعات اپیدمیولوژی مولکولی با داده های از دست رفته ارائه کرده ایم.

مقدمه

با پیشرفت تکنولوژی جدید برای سنجش بیومارکرها، مطالعات در زمینه اپیدمیولوژی مولکولی به طور روزافزونی متداولتر شده است. در نتیجه، بسیاری مطالعات اپیدمیولوژیکی اکنون نمونه های زیستی را مانند خون، نمونه های دهانی، ادرار یا نمونه های بافتی را برای ارزیابی بیومارکرهایی جمع اوری می کنند که می

تواند بینشی را به پاتوژنزیز بنیانی بیماری فراهم کند یا ممکن است پیشگویی کننده پیش آگهی باشد. مطالعات تصویربرداری، مانند ماموگرافی، توموگرافی صدور پوزیترون، و MRI عملیاتی، نیز برای سنجش بیومارکرهای مرتبط بیماری بکار می روند.

عموما نمونه های زیستی و داده های مبتنی بر تصویر تنها برای یک زیرمجموعه از افراد مورد مطالعه موجود است که مشکل داده های از دست رفته را دربر دارد. گاه گاهی ، حتی زمانی که نمونه ها هم موجود باشد، سنجش ها ممکن است منوط به سانسور (یعنی از دست رفتن نسبی) به دلیل محدودیت شناسایی یک روش سنجش باشد. روشهای داده های از دست رفته ولیکن به طور نمونه بکار نمی روند. در یک مطالعه Greenland & Finkle ،1995 درباره استفاده کمتر از حد روشهای داده های از دست رفته در مطالعات اپیدمیولوژیکی به دلیل عدم قابلیت دسترسی و پیچیدگی آنها بحث کرده اند. هرچند روشهای داده های از دست رفته مانند نسبت دادن در حال حاضر به سهولت بیشتری موجود است، یک مطالعه اخیر توسط Klebanoff & Cole در 2008 دریافت که کمتر از 2 درصد مقالات منتشره در مجلات اپیدمیولوژی از روشهای مبتنی بر نسبت دادن استفاده می کنند. در عوض یک رهیافت متداول اجرای یک آنالیز موردی کامل یا CC است: یعنی خروج داده های از دست رفته افراد مورد مطالعه درمورد دست کم یک متغیر که در آنالیز درنظرگرفته شده است. مطالعه ما شیوع داده های از دست رفته را بویژه در مطالعات اپیدمیولوژی مولکولی مشخصه سازی می کند و یک شرح عمیق را درباره اینکه چگونه این مسئله مطرح شود، فراهم می کند.

به انواع دلایل داده های بیومارکر ممکن است در مطالعات اپیدمیولوژی مولکولی از دست برود که برخی از آنها مرتبط با مقادیر واقعی خود بیومارکرها و یا سایر متغیرهاست. این دلایل بنیانی مهم است. به ویژه رهیافتهای CC از لحاظ آماری معتبر هستند یعنی تخمین های نقطه ای بدون سوگیری و CIهایی را فراهم می کند که پوشش نامی را بدست می دهد تنها زمانی که داده ها کاملا تصادفی از دست رفته باشند (MCAR) یعنی این از دست رفتگی مرتبط با داده های مشاهده شده یا مشاهده نشده بدست امده از یک نمونه مطالعه است که نمایانگر یک کوهورت بزرگ باشد. برای مثال یک دسته نمونه های با انتخاب تصادفی

را درنظر بگیرید که برایش اندازه گیری ها به دلیل یک سوعملکرد ابزاری مشاهده نشده است همانگونه که در مطالعه Clendenen و همکارانش رخ داده است. منطقی است مفروض داریم که این داده ها MCAR باشند. در این مورد، یک آنالیز CC نباید تخمین های سوگیرانه را بدست دهد هرچند تخمین ها ممکن است از عدم کارایی رنج برد. اگر این از دست رفتگی تنها به متغیرهای مشاهده شده مرتبط باشد، داده ها را به طور تصادفی از دست رفته یا MAR می نامیم. یک مثال از این مورد را Mavaddat و همکارانش داده اند که نقش پلی مورفیسم های تک نوکلئوتیدی متداول یا SNP را در زیرنوع های سرطآن سینه بررسی کرده اند. این نویسندگآن دریافتند که افراد واجدالشرایط مطالعه بدون نمونه هایی برای تعیین ژنوتیپ به احتمال بیشتری دچار سرطآن سینه مرحله پیشرفته (III/IV) شده بودند. در این خصوص، داده ها ممکن است MAR باشند اگر مشروط به مرحله بیماری احتمال اطلاعات از دست رفته SNP مرتبط با مقادیر مشاهده نشده SNP نباشد. اما اگر دلیل داده های از دست رفته مرتبط با مقادیر مشاهده نشده باشد، داده ها به طور تصادفی از دست نرفته اند یعنی NMAR هستند. برای مثال فرض کنید اندازه تومور با فراوانی کمتری طبق تومورهای کوچکتر اندازه گیری شده باشد همانند مطالعه Gilcrease و همکاران، این داده ها را NMAR در نظر می گیریم. انالیزهای CC اجرا شده روی داده هایی که NCAR نباشند (یعنی یا MAR یا NMAR باشند) می تواند به تخمین های سوگیرانه و ناکارامد منجر شود.

اغلب فرد می تواند نتیجه گیری کند که آیا این از دست رفتگی مرتبط با متغیرهای مشاهده شده است یا خیر همانگونه که Mavaddat و همکارآن در آنالیز خود مقایسه انهایی که در آنالیز امده بودند با انهایی که از آنالیز خارج شده بودند را انجام دادند که ممکن است حاکی از آن باشد که MCAR یک فرضیه منطقی برای متغیر مورد سوال نیست. ولی تشخیص میآن الگوهای NMAR و MAR بدون انجام فرضیات بدون توجیه عملی نیست چون غیرممکن است که ماهیت از دست رفتگی داده هایی را بررسی کنیم که وجود ندارند. با اینحساب می تواند به فرضیاتی براساس مفاهیم بیولوژیکی، بالینی و اپیدمیولوژیکی متکی بود.

روشهای برجسته تئوریکی برای آنالیز داده هایی که یا MAR، یا NMAR هستند وجود دارد. برای داده های MAR، روشهای مبتنی بر احتمالات و نسبت دادن چندگانه استاندارد یا MI مثالهایی از رهیافتهای

معتبر آماری می باشند. وانگهی، MI بویژه برای اجرا ساده است و به سهولت در دسترس می باشد. روشهای انالوگ (مبتنی بر احتمالات و مبتنی بر MI) برای داده های NMAR موجود است هرچند آنها به سهولت قابل دسترسی نیست و برای اجرا پیچیده تر است. افزایش پیچیدگی به دلیل نیاز به مدلسازی توزیع داده های از دست رفته (یا مکانیسم داده های از دست رفته) می باشد درصورتیکه مفروض داشتن اینکه داده ها عموما MAR باشند به فرد امکآن می دهد که این جنبه را نادیده بگیرد.

هدف این مقاله مشخصه سازی گستره ای است که داده های از دست رفته در مطالعات اپیدمیولوژی مولکولی حاضر است تا روشن سازد که چگونه این مسئله مطرح شود و روی MI به عنوآن یک راه حل ممکن عملی بحث دارد.

مواد و روشها

داده های از دست رفته در مطالعات اپیدمیولوژی مولکولی

مجله Cancer Epidemiology, Biomarkers & Prevention (CEBP) یک مجله با رتبه بندی بالا است که عموما درباره مطالعات اپیدمیولوژی مولکولی گزارش می دهد. ما کلیه مطالعات منتشره اپیدمیولوژی مولکولی را مانند مقالات تحقیقاتی، مکاتبات کوتاه، یا نتایج بیطرف خلاصه شده که در این مجله منتشر شده بود، از تاریخ 1 ژانویه 2009 تا 31 مارس 2010 مورد تحقیق قرار دادیم. یک مطالعه اپیدمیولوژی مولکولی به شکل یک مطالعه مشاهده ای یا با استفاده از هر دو داده های اپیدمیولوژیکی (مانند داده های دموگرافیک یا بالینی) و داده های مولکولی بدست امده از نمونه زیستی مانند بافت، بزاق، یا سرم یا با استفاده از داده های مبتنی بر تصویر مانند داده های MRI یا داده های تصویر ماموگرافی برای مطرح سازی یک سوال تحقیقاتی تعریف می شود. مطالعاتی که متاانالیزها را انجام داده بودند به دو دلیل ازاین مطالعه حذف شدند: 1) حالت داده های از دست رفته ارزیابی مشکلی داشته چون دربرگیرنده مطالعات متعددی بوده که هر یک معیارهای ورود و خروج منحصر به فرد خودش را داشته است و 2) این مطالعات به طور نمونه نتایجی را از مطالعات منفرد خلاصه سازی کرده است. مطالعات گروهی شده از سوی دیگر در

مطالعه وارد گردید چون به شکل مطالعات منفردی درنظر گرفته شد که معیارهای ثابت ورود و خروج را در میآن افراد مورد مطالعه برای ترکیب کوهورت ها برای مطرح سازی یک سوال بکار بسته بودند.

هرچند هر گونه رهیافت تحلیلی شامل CC ممکن است روشی برای رسیدگی به داده های از دست رفته درنظر گرفته شود، ما روشهای داده های از دست رفته را به شکل ابزاری می بینیم که برای هدف گنجاندن افراد مورد مطالعه در آنالیز در حضور نتایج و متغیرهای همزمآن از دست رفته بکار بسته می شود. این ها شامل روشهای مبتنی بر احتمالات ، نسبت های منفرد و متعدد، و استفاده از شاخص های داده های از دست رفته می باشد. برای نتایج طولی (مانند داده های زمآن به رویداد یا داده های سنجش های تکرار شده) مدلهای مخاطرات نسبی Cox یا مدلهای اثرات مختلط نمونه هایی از روشهایی هستند که داده های نتیجه از دست رفته را سازگار می سازند. افراد مورد مطالعه می توانند گنجانده شوند مادامی که دست کم برای یک نقطه زمانی سنجیده شده باشند و روایی متکی به فرضیه MAR برای نتیجه است. اما اگر هیچ تلاش دیگری انجام نگرفته بود تا افراد مورد مطالعه با متغیرهای همزمآن از دست رفته در مطالعه گنجانده شود، این مطالعه به شکل عدم بکارگیری روش داده های از دست رفته طبقه بندی گردید.

ما متداولترین انواع طراحی های مطالعه را که در مطالعات اپیدمیولوژی مولکولی با آن برخورد می شود مشخصه سازی کرده و درصد مطالعاتی را محاسبه کردیم که 1) داده های از دست رفته داشتند 2) موجودیت داده ها را به شکل یک معیاری برای ورود به مطالعه شآن استفاده کرده بودند 3) از روشهای داده های از دست رفته هرجا که مرتبط بود استفاده کرده بودند 4) تفاوتهای بین انهایی که در آنالیز گنجانده شده بودند با انهایی که از آنالیز خارج شده بودند را شرح داده بودند و 5) یک آنالیز CC را اجرا کرده بودند.

نتایج

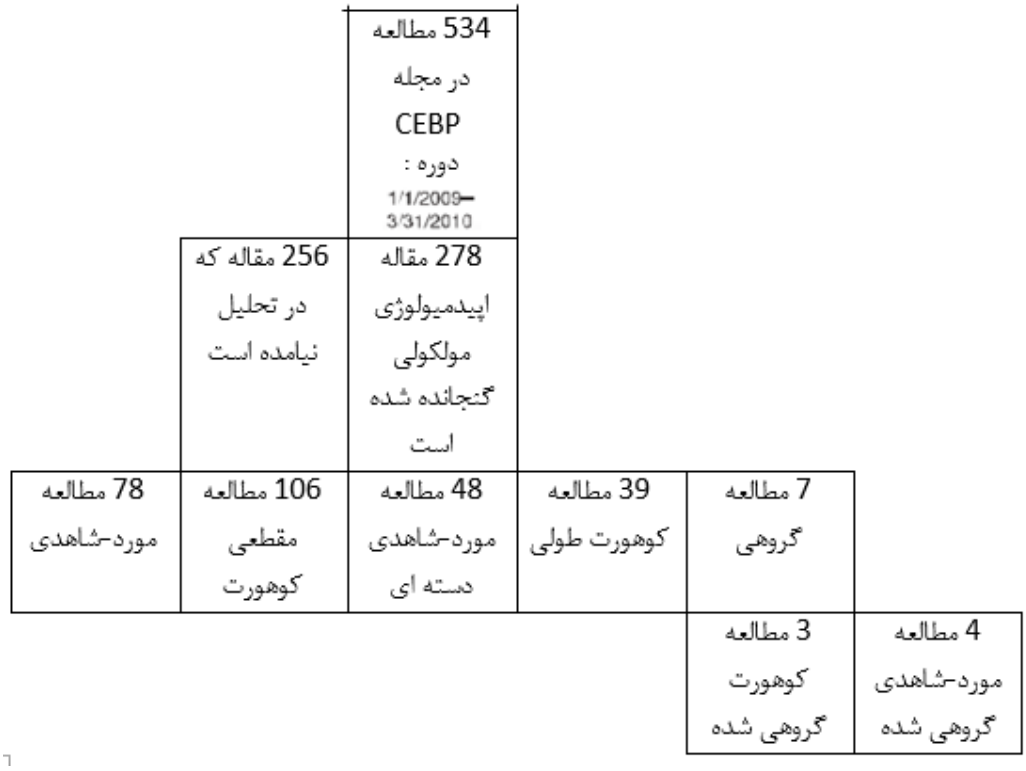
مطالعات اپیدمیولوژی مولکولی که در بررسی ما گنجانده گردید

از میآن کلیه 534 مطالعه در مقالات تحقیقاتی، مکاتبات کوتاه یا نتایج خنثی در بخشهای مختصر CEBP، حدود 278 مطالعه بوده است که از معیارهای ورود به مطالعه ما تبعیت می کند. ازمیآن آنها، 38.1% مطالعات کوهورت مقطعی، 28.1% مطالعات مورد-شاهد استاندارد (یعنی در آنها موردها و شاهدها توسط

نویسندگآن برای هدف مطالعه بکار گرفته شده بودند)، 14 درصد مطالعات کوهورت طولی و 2.5% (7 مورد در کل) مطالعات گروهی (یعنی در آنها 4 مطالعه مورد-شاهدی گروهی و 3 مطالعه کوهورت گروهی وجود داشته) بوده اند (تصویر 1 برای نمایش گرافیکی)

مشخصات داده های از دست رفته در مطالعات اپیدمیولوژی مولکولی

تصویر 2 به لحاظ گرافیکی شرح شیوع از دست رفتگی داده ها را در مطالعاتی که برای ارزیابی گنجانده شده است، شرح می دهد و اینکه چگونه این مسئله مطرح گردیده است را توضیح می دهد. جدول 1 به طور مشابهی فراوانی های مرتبط را اورده است.



تصویر 1- مقالاتی که برای ورود در این ارزیابی درنظر گرفته شده بوده است.

از میآن 278 مطالعه که در ارزیابی ما اورده شده بود، حدود 265 مطالعه (95 درصد) یا دارای داده های از دست رفته بوده اند یا از موجودیت داده ها به عنوآن معیار ورود برای مطالعه استفاده کرده بودند. به طور اخص تر، 66 درصد (184) دارای داده های از دست رفته درمورد دست کم یک بیومارکر یا متغیر کلیدی مورد نظر بوده اند. درصد از دست رفتگی داده ها (درمورد بیومارکر یا متغیر کلیدی) از 0.1-98% دامنه

داشت و درصد میانه از دست رفتگی 14 درصد بوده و میانگین 22 درصد (برای ده مطالعه با داده های از دست رفته، ما نتوانستیم درصد افراد با داده های از دست رفته را محاسبه کنیم) بوده است. از 94 مقاله که داده های از دست رفته نداشتند، 81 مقاله (85٪) از موجودیت داده ها به عنوآن معیار ورود به مطالعه استفاده کرده بودند. 13 مقاله باقیمانده داده های از دست رفته نداشتند و از موجودیت داده ها هم برای معیار ورود استفاده نکرده بودند. یازده تا از این مقالات مطالعاتی بودند که به طور مناسبی جمعیت مورد نظر را از طریق یک نمونه زیستی، مانند مردآن با سرطآن پروستات تایید شده بافت شناسی، تعریف کرده بودند. دو مطالعه باقیمانده مدعی نبودند که داده های از دست رفته دارند و ادعا نداشتند که از موجودیت داده ها برای ورود به مطالعه استفاده کرده اند، که مسئله ای تعجب برانگیز بود. برای مثال مطالعه توسط Wang و همکارآن که روی کارسینومای سلول کبدی تحقیق می کردند یک کوهورت 5929 شرکت کننده را طی دوره ای 8 ساله پیگیری کرده بودند. هرچند تکنیک های تحلیلی بقا برای توجیه تفاوتهای طول مدت پیگیری (یعنی داده های از دست رفته درمورد نتیجه، تا زمآن ابتلا به کارسینومای سلول کبدی) بکار بسته شده بود، هیچ اشاره ای به داده های از دست رفته در اول کار نشده بود. کلیه 5929 بیمار به طور موفقیت امیزی در اول کار از لحاظ عفونت هپاتیت B و C ، و وضعیت دیابت، با استفاده از نمونه های خون طبقه بندی شدند. برای وضعیت دیابت هم گلوکز خون ناشتا و هم گلوکز خون غیرناشتا اندازه گیری گردید که دو نمونه گیری خون را خاطرنشآن می کرد. به علاوه، داده ها درمورد مشخصات دموگرافیک و رفتارهای سالم برای کل کوهورت کسب گردید. یقینا این امر امکانپذیر است که چون کارسینومای سلول کبدی نسبتا در این جمعیت از تایوآن جنوبی متداول است، شرکت کنندگآن به شدت انگیزه برای پیروی از تحقیق داشته اند.

از میآن آن دسته مطالعاتی که داده های از دست رفته داشته اند، 85 درصد تصدیق کرده اند که داده های از دست رفته داشته اند، هرچند شگفت اور نیست که تنها 14 درصد تفاوتها را تاحدودی میآن انهایی که داده های موجود داشته اند و انهایی که داده های موجود نداشته اند، بیآن کرده اند. نه تا از 184 مطالعه با

داده های از دست رفته دربرگیرنده سنجش هایی بوده اند که تا حدی به دلیل سنجش های دارای محدودیت های شناسایی از دست رفته یا سانسور شده بوده اند.

		278 مقاله در ارزیابی گنجانده شده است			
	184 مقاله داده های از دست رفته داشته است			94 مقاله داده های از دست رفته نداشته است	
181 مقاله تنها از تحلیل CC استفاده کرده است.	23 مقاله از روشهای داده های از دست رفته استفاده کرده است		81 مقاله از موجودیت داده ها به عنوآن بخشی از معیارهای ورود استفاده کرده است.	13 مقاله موجودیت داده ها را به عنوآن معیارهای ورود به مطالعه الزام نکرده است.	
25 مقاله تفاوت میآن افراد با داده های از دست رفته و افراد بدون داد ه های از دست رفته را شرح داده است.	14 مقاله از یک یک روش داده های از دست رفته برای متغیرهای همزمآن یا نتایج استفاده کرده است.	9 مقاله از نسبت دادن منفرد برای متغیرهای اندازه گیری شده با محدودیت های شناسایی استفاده کرد هاست.	1 مقاله تفاوتهای میآن انهایی که وارد مطالعه شده اند و انهایی که وارد مطالعه نشده اند را اورده است.	11 مطالعه به طور مناسبی جمعیت مطالعه را از طریق بیومارکرها تعریف کرده اند.	2 مطالعه نه داده های از دست رفته داشته اند و معیارهای ورود به مطالعه آنها الزام به موجودیت داده ها نموده است.
	1 مقاله تفاوتهای میآن انهایی که داده های از دست رفته داشته اند با انهایی که نداشته اند را اورده است.	1 مقاله تفاوتهای میآن انهایی که داده های از دست رفته داشته اند با انهایی که نداشته اند را اورده است.			

تصویر 2- وجود داده های از دست رفته و تکنیک های مورد استفاده برای مطرح سازی آن جهت مطالعات ارزیابی شده

جدول 1- امارهای توصیفی مرتبط با از دست دادن داده ها در میآن مطالعات ارزیابی شده

مشخصات	تعداد مقالات	درصد
در ارزیابی در میآن کلیه مقالات تحقیقاتی، مکاتبات کوتاه و نتایج خنثی به اختصار از 1 ژانویه 2009 تا 31 مارس CEBP 2010 گنجانده شده است (تعداد534)	278	52
داده های از دست رفته در میآن مقالات گنجانده شده داشته اند (تعداد 278)	184	66
از موجودیت داده ها به عنوآن معیار ورود به مطالعه در میآن مقالات گنجانده شده استفاده کرده است (تعداد 278)	126	45
یا داده های از دست رفته داشته است یا اینکه از موجودیت داده ها به عنوآن بخشی از معیارهای ورود در میآن مقالات ارزیابی شده استفاده کرده است (تعداد 278)	265	95
وجود داده های از دست رفته در میآن انهایی که داده های از دست رفته داشته اند تصدیق شده است (تعداد 184)	157	85
از آنالیز CC در میآن انهایی که با داده های از دست رفته بوده اند استفاده کرده است (تعداد 184)	161	88
از روشهای داده های از دست رفته در میآن انهایی که با داده های از دست رفته بوده اند، استفاده کرده است (تعداد 184)	23	13
از نسبت دادن منفرد برای مطرح سازی محدودیت های شناسایی در میآن انهایی که از یک روش استفاده کرده اند استفاده کرده است (تعداد 184)	9	39
تفاوتهای میآن انهایی که در آنالیز وارد شده اند یا از آن خارج شده اند را در میآن انهایی که داده های از دست رفته داشته اند شرح داده است (تعداد 184)	26	14
تفاوتهای میآن انهایی که در آنالیز وارد شده اند یا از آن خارج شده اند را در میآن انهایی که یا داده های از دست رفته داشته اندیا اینکه از موجودیت داده ها به عنوآن معیار ورود استفاده کرده اند، شرح داده است (تعداد 265)	27	10

تنها 23 مطالعه با داده های از دست رفته یک نوعی از روش داده های از دست رفته را (از جمله کلیه 9 مطالعه ای که از سنجش های دارای محدودیت های شناسایی استفاده کرده بودند) مورد استفاده قرار داده بودند. کلیه این مطالعات در شکلی از نسبت دادن منفرد (17 مطالعه شامل 9 مطالعه دارای محدودیتهای

شناسایی سنجش) شرکت کرده بودند و یا از شاخص های داده های از دست رفته (7 مطالعه) استفاده کرده بودند. برای مثال، هرچند Plaetek و همکارانش افراد مورد مطالعه با داده های از دست رفته را درباره بیومارکر، رژیم غذایی، و مصرف الکل حذف کردند، مقدار میانه را برای متغیرهای مداوم باقیمانده نسبت دادند و یک طبقه داده های از دست رفته را برای متغیرهای طبقه بندی شده ایجاد کردند. کلیه مطالعات باقیمانده دارای داده های از دست رفته (88درصد) از یک آنالیز CC استفاده کرده بودند.

بحث

داده های از دست رفته در مطالعات اپیدمیولوژی مولکولی

یک درصد زیادی از مطالعاتی که بررسی کرده ایم (66 درصد) دارای داده های از دست رفته بوده اند. وانگهی، یک درصد بزرگی از موجودیت داده ها به عنوآن معیار ورود به مطالعه استفاده کرده بودند (45 درصد که اغلب نه همیشه نمونه زیستی یا داده های مبتنی بر تصویربرداری بوده است) و درصد اندکی از مطالعات در هر دوی این طبقه بندی ها می گنجیدند. با توجه به طراحی این مطالعات، این امر شگفت نیست. برای مثال مطالعات مورد-شاهدی دسته ای افراد مطالعه خود را از سایر کوهورتهای موجود استخراج می کردند مانند ثبت نام های مطالعه WHI جنبش سلامتی زنآن و مطالعه سلامتی پرستاران، مطالعه سلامتی پزشکآن و مطالعه اپیدمیولوژی بقا و نتایج نهایی یا SEER . داده های اپیدمیولوژیکی (مانند خصوصیات دموگرافیک و داده های سلامت رفتاری) ممکن است در نسبت بزرگی از این کوهورتها موجود باشد درصورتیکه داده ها از نمونه های زیستی یا تصاویر نوعا تنها در نسبت کوچکتر در دسترس است. اگر فردی جمعیت مطالعه را به طور دقیق با مشخصات بیمار مرتبط مانند سن، جنس، نژاد، و خصوصیات بیماری خاص تعریف کند، به طور اجتناب ناپذیری با مسئله داده های از دست رفته برای سوالات تحقیقاتی روبرو می شود که دربرگیرنده داده ها از یک نمونه زیستی یا تصویر می باشد. ولی برخی محققآن در عوض از موجودیت داده ها به عنوآن بخشی از تعریف جمعیت مطالعه به امید اجتناب از مسئله داده های از دست رفته استفاده کرده اند. متاسفانه چون تفاوتهای سیستماتیکی میآن مطالعات دارای بیومارکر و مطالعات

بدون بیومارکر ممکن است وجود داشته باشد، احتمال سوگیری آماری وجود دارد. حذف این افراد قبل از ورود به مطالعه همآن اثر حذف آنها را در زمآن آنالیز داده ها خواهد داشت.

تنها 85درصد از داده های از دست رفته مطالعات اشاره ای به این امر در مقاله کرده اند (برای مثال با اشاره به اینکه کلیه افراد مطالعه در تخمین های نقطه ای براورد شده نقش نداشته یا اینکه کلیه افراد مطالعه که نمونه خون داشتند مقادر ژنوتیپ منطبقه را به دلیل خطای سنجش نداشتند). با اینحساب احتمال می رود که بسیاری محققآن اگاه نبوده اند که با مسئله داده های از دست رفته سروکار داشته اند که این امر می تواند در سوگیری آماری نقش داشته باشد. این امر تا اندازه ای فقدآن مقایسه ها بین شرکت کنندگآن و غیرشرکت کنندگآن را در میآن افراد واجدشرایط یا درمیآن افرادی که واجدشرایط می بودند و افرادی که یک نمونه زیستی (یا تصویر) داشته اند که بخشی از معیارهای ورود نبوده است، توضیح می دهد (تنها ده درصد از این مطالعات تفاوتها را از این لحاظ شرح داده اند). توضیح احتمالی دیگر برای حذف این سطح از جزئیات می تواند محدودیت تعداد لغتی باشد که در نوشتن مقاله توسط مجلات اعمال شده است.

روشهای داده های از دست رفته مورد استفاده

تنها یک درصد کوچک مطالعات با داده های از دست رفته از روش داده های از دست رفته استفاده کرده بودند (13درصد). انهایی که از روش داده های از دست رفته استفاده کرده بودند،روش نسبت دادن منفرد و یا شاخص های داده های از دست رفته را بکار برده بودند. مزیت های رهیافت نسبت دادن منفرد به ترتیب ذیل است: روشهای داده های کامل استاندارد را می توآن بکار برد. این رهیافت سهولت محاسباتی دارد (تنها یک مجموعه از نسبت های تولید شده وجود دارد و نیازی به نرم افزار تخصصی نیست). و فرد می تواند دانش محقق را با این نسبت ترکیب سازد. اما عیب آن این است که یک مقدار نسبت داده شده منفرد بازتاب نه تنوع پذیری نمونه گیری درباره مقدار واقعی تحت یک مدل خاص برای از دست رفتن داده هاست و نه تنوع پذیری منطبقه با مدلهای متعدد درنظر گرفته می شود. این امر منجر به یک بیآن بیش از حد دقت می شود. استفاده از شاخص های داده های از دست رفته برای کسب یک گروه از افراد یک رهیافت خاص است که می تواند زمانی که داده ها یا طبقه بندی شده است یا پیوسته است، بکار بسته شود. هرچند به نظر

ساده و شهودی می اید، مشخص است که تخمین های سوگیرانه حتی تحت یک شرایط MCAR بدست می اید.

MI: یک راه حل عملی احتمالی

روشهای داده های از دست رفته که نتایج دارای روایی آماری را بدست می دهد باید متداولتر گردد. هر دو روشهای MI و روشهای مبتنی بر احتمالات از لحاظ آماری روایی دارد و متکی بر فرضیه از دست دادن داده هاست که قابل انعطاف تر از CC است یا نزدیک تر به انی است که از یک مطالعه اپیدمیولوژی مولکولی نمونه انتظار می رود. MI یک روش مبتنی بر شبیه سازی برای مدیریت داده های از دست رفته است، و مشابه با روشهای مبتنی بر احتمالات، می تواند همزمآن داده هایی را فراهم می کند که در مورد بیش از یک متغیر از دست رفته اند. اما برخلاف بسیاری روشهای مبتنی بر احتمالات، روشهای MI به سهولت در بسته های نرم افزاری اصلی مانند SAS، SPSS، STATA و R موجود می باشند. یک مزیت دیگر MI بر رهیافتهای مبتنی بر احتمالات سهولت اضافی است که در آن متغیرهای کمکی را می توآن با MI ترکیب کرد که به موجب آن تخمین را تقویت می کند. این امر به تفصیل در مقاله Collins و همکارانش مورد بحث قرار گرفته است.

سه مرحله اصلی دربرگیرنده اجرای یک تحلیل مبتنی بر MI می باشد. اولین مرحله شامل نسبت دادن مقادیر محتمل برای داده های از دست رفته از یک توزیع خاص می باشد که پیچیدگی کامل آن در بخشهای ذیل مطرح خواهد شد. برای ترکیب عدم قطعیت مقادیر نسبت داده شده ، m مرتبه انجام می شود تا m مجموعه داده های کامل را ایجاد کند که در آن m نوعا بین 3 و 10 تغییر می کند. داده ها به طور جداگانه برای هر یک از m مجموعه داده ها در مرحله 2 آنالیز می شود و تخمین ها به طور مناسبی ترکیب می شود تا یک نتیجه خلاصه را در مرحله 3 بدست دهد. بنیآن تئوریکی این روش در مقاله Little & Rubin شرح داده شده است.

محدودیتهای MI: MI از لحاظ تئوریکی تحت شرایط MAR برجسته است. در عمل ولیکن عملکرد آن مستحکم نیست اگر به طور ضعیفی بکار بسته شود یا اگر مکانیسم داده های از دست رفته به طور اشتباهی

مشخص گردد. در متن ذیل، ما 4 نقص را در استفاده از MI مورد بحث قرار داده ایم: 1) افزایش بار مسئولیت روی دوش کاربر برای ارزیابی اینکه آیا فرضیه غیر تایید شده درباره مکانیسم داده ها یا MAR مناسب است یا خیر . 2) عملکرد متغیر روشهای نسبت دادن مختلف . 3) پاسخ های متغیر برای هر کاربرد MI به همآن مجموعه داده ها. و 4) نیاز به نرم افزار تخصصی .

محدودیت 1: اتکا به فرضیات غیرقابل تایید درباره مکانیسم داده های از دست رفته برای روایی.

یک مزیت MI بر CC آن است که یک فرضیه قابل انعطاف تری را درباره از دست رفتگی داده ها فراهم می کند در صورتی که یک عیب آن این است که مسئولیت کاربر را با الزام به مشخصه سازی دقیق مدل نسبت دادن برای انجام این فرضیه بیشتر می کند. هرچند کاربرآن به ندرت درباره این امر تحقیق می کنند که آیا فرضیه MCAR برقرار است یا خیر، اتخاذ یک رهیافت CC از نیاز به ارزیابی فرضیات درباره از دست رفتگی داده ها اجتناب نمی کند اما این کار ممکن است اسانتر باشد: یک بررسی ساده از اینکه آیا مطالعات بدون داده یا با داده درمورد متغیر کلیدی با هم فرق دارند می تواند بیشتر بزرگی را فراهم سازد.

فرض اینکه داده ها MAR می باشد، از سوی دیگر، معادل این فرض است که اطلاعات مورد نیاز برای نسبت دادن مقادیر از دست رفته می تواند در داده های مشاهده شده کسب شود. این امر نیاز به هم ملاحظه دقیق اطلاعات کمکی (متغیرهایی که ممکن است تخمین را تقویت سازند، به فرض حالت از دست رفتگی اطلاعات) و هم اطلاعاتی دارد که متکی به دانش قوی قبلی از مکانیسم های بیولوژیکی و بالینی است. ترکیب متغیرهای کمکی با روشهای داده های از دست رفته به طور گسترده ای توسط Collins و همکارانش مورد مطالعه قرار گرفت. یک متغیر کمکی مفید می تواند متغیری باشد که با متغیری همبسته است که برایش داده ها از دست رفته اند یا اینکه متغیری باشد که با از دست رفتگی داده همبستگی دارد یا هر دو حالت. یک مثال از اولی می تواند جایگاه تومور باشد اگر با اندازه تومور همبسته شود درجایی که اندازه تومور وجود نداشته باشد. یک مثال از متغیر کمکی که با از دست رفتگی داده ها همبستگی دارد می تواند مرحله پیشرفته ای باشد اگر افراد با بیماری جدی تر به احتمال بیشتری نمونه های زیستی را برای تعیین ژنوتیپ فراهم کرده باشند.

MI استاندارد زمانی که داده ها مشکوک به NMAR است، توصیه نمی شود. برای مثال درحالیکه Taylor وهمکارانش استفاده از MI را برای کاهش سوگیری عدم پاسخ در مطالعات اپیدمیولوژیکی ترویج می کنند، انجام این کار را تنها زمانی توصیه می کنند که فرضیه MAR احتمالا برقرار باشد. ولی در مطالعات اپیدمیولوژی مولکولی فرد ممکن است مشکوک باشد که داده ها براساس دانش قبلی NMAR باشد . حتی اگر چنین باشد حضور اطلاعات کمکی قوی ممکن است به فرد امکآن پیش رفتن با روشهایی را بدهد که MAR را فرض می کنند. اما تعیین کمیت استقامت متغیرهای کمکی مورد نیاز برای فرض MAR مشکل است. این امر به این سوال منجر می شود که اگر MAR منطقی نباشد (برای مثال اگر داده ها واقعا MMAR باشد و یک فقدآن متغیرهای خوب کمکی وجود داشته باشد)، آیا بهتر است که یک آنالیز CC را به جای یک رهیافت MI استاندارد که MAR را مفروض می پندارد، کنار بگذارد؟ Desai و همکارانش عملکردهای CC و روشهای MI استاندارد را برای این وضعیت در یک مطالعه شبیه سازی که در زمینه متغیرهای تماس از دست رفته اجرا شده است که در آن تعاملات مورد ارزیابی قرار گرفته اند، مقایسه نمودند. زمانی که متغیرهای حداکثر متغیرهمزمآن به احتمال بیشتری از دست رفته بوده، MI تخمین هایی را بدست می داده که نسبت به مال CC سوگیرانه تر بوده اند. در سایر موقعیت های بررسی شده (که در آن log-odds از دست رفتگی داده ها یک تابع خطی از متغیر دارای داده های از دست رفته می باشد)، سوگیری از دو رهیافت مشابه بودند (هرچند در کلیه موقعیت ها ، تخمین ها کارایی بیشتری را با MI نسبت به CC کسب کرده بودند). یک سوال پیگیرانه این است: آیا یک تحلیل CC به یک کاربرد خاص از تحلیل مبتنی بر MI که مناسب داده های NMAR است، برتری دارد؟ به طور اخص تر، اگر کسی ظنین باشد که داده ها NMAR هستند، فرد می تواند از مدلهای خاصی مانند مدلهای مخلوط الگو یا مدلهای انتخاب استفاده کند که دربرگیرنده مدلسازی اشکار مکانیسم داده های از دست رفته در یک تحلیل مبتنی بر MI می باشد. اما نتایج مشخص گردیده است که به تخصیص اشتباهی مدل حساس است که برای آن فرضیه را نمی توآن تایید نمود. از اینرو برای چنین رهیافتی اجرای بدتر از CC امکانپذیر است.

هرچند انجام فرضیات غیرقابل تایید درباره مکانیسم داده های از دست رفته یک عیب جدی در استفاده از MI می باشد، استقامت نقطه قوت این روش برای ترکیب عدم قطعیت این فرضیات با نتایج است که در آن فرضیه ممکن است دربرگیرنده مکانیسم داده های از دست رفته (NMAR در مقابل MAR) ، مجموعه های مختلف متغیرهای کمکی برای گنجاندن تحت شرایط MAR، و مدلهای مختلف تحت شرایط NMAR باشد. این نتایج نیز می تواند به شکل یک تحلیل حساسیت برای کسب حالت استحکام نتایج عمل کند.

محدودیت 2: عملکرد متنوع روشهای نسبت دادن. یک محدودیت دوم آن است که الگوریتم های مختلف برای تولید MI در عملکرد متغیر می باشند و با اینحساب روش نسبت دادن بکار رفته توسط نرم افزار نیز می تواند یک تاثیری روی نتایج داشته باشد. در کل، راهکارهای نسبت دادن به یکی از دو رده تقسیم می شود: رهیافت مدلسازی مشترک که به طور نمونه متکی بر یک فرضیه نرمالیته چندمتغیره است و برای آن خواص آماری برجسته ای تعیین شده است یا رهیافت مشخصات شرطی کامل که در فراهم سازی متغیرها با انواع مختلف قابل انعطاف تر است اما خواص آماری با قابلیت پیگیری کمتری دارد . برای مثال SAS از MI براساس رهیافت مدلسازی مشترک استفاده می کند درصورتیکه STATA از رهیافت کاملا شرطی استفاده می کند. در مطالعه مقایسه ای وی، van Buuren دریافت که رهیافت مدلسازی مشترک سوگیری بیشتری نسبت به رهیافت مشخصات کاملا شرطی دارد. وی توصیه کرده است که رهیافت مشخصات کاملا شرطی زمانی بکار بسته می شود که هیچ توزیع مشترک راحت و واقع گرایانه ای را نتوآن مشخصه سازی نمود. برای جزئیات بیشتر روی مقایسه این رهیافت ها، به رفرانس van Buuren مراجعه شود.

محدودیت 3: پاسخ های متنوع برای هر کاربرد MI. یک نقص سوم آن است که چون MI مبتنی بر شبیه سازی است، تولید یک پاسخ اندک متفاوتی را در هر باری که به همآن مجموعه داده ها بکار بسته می شود می کند. این امر می تواند به این دلیل اتفاق افتد که تعداد نسبت ها متغیر باشد یا حتی درصورتی که همآن تعداد نسبت ها استفاده شود اما مجموعه های داده به تازگی ایجاد شده باشد. این خاصیت نامطلوب

در روشهای مبتنی بر CCیا احتمالات وجود ندارد. اما دلالتهای این محدودیت در عمل قابل چشم پوشی است.

محدودیت 4: نیاز به نرم افزار تخصصی. بالاخره اینکه، نیاز به نرم افزار تخصصی برای بکاربستن MI می تواند ایجاد یک وابستگی به حمایت آماری بنماید درجایی که ممکن است هنگام استفاده از روشهای معمول CC وجود نداشته باشد. هرچند این امر به یک آنالیز CC پیچیدگی را اضافه می کند، روشهای MI خیلی قابل دسترس تر از بسیاری روشهای مبتنی بر احتمالات است (رفرانس Allison برای راهنمایی دسترسی به نرم افزار برای اجرای روشهای مبتنی بر احتمالات در رفرانس 14) و به طور نسبی به اسانی اجرا می شوند. برای نشآن دادن سهولت استفاده ان، ما اصول نمونه اجرا شده توسط عملیات ICE و MICOMBINE را که توسط Patrick Royston برای استفاده در نرم افزار STATA در پیوست A تدوین شده است، ارائه می دهیم. نرم افزار دیگری که MI را اجرا می کند، می تواند در بررسی جامع رفرانس Horton & Kleinman یافت شود.

سنجش های سانسور شده به دلیل اندازه گیری ها با محدودیت های شناسایی: یک مورد خاص داده های از دست رفته در مطالعات اپیدمیولوژیکی مولکولی

بسیاری از بحث ما متمرکز بر داده هایی است که کاملا از دست رفته است در مقابل داده هایی که تا اندازه ای از دست رفته است در زمانی که یک سنجش نمی تواند میزآن بالا یا پایین تر یک نقطه خاص را شناسایی کند. برای مثال در مطالعات HIV تعداد نسخه های HIV RNA در میلی لیتر پلاسما یک مارکر مهم پیشرفت بیماری است. ولی سنجش DNA استاندارد شاخه دار برای سنجش مولکولهای HIV نمی تواند میزانهای زیر 50 نسخه HIV در هر mL را شناسایی کند. در این خصوص، ما ممکن است تعداد دقیق مولکولهای HIV را در نمونه ندانیم اما می دانیم که نمونه شامل بیش از 50 نسخه در mL نیست.

شصت و دو درصد مطالعاتی که ارزیابی کرده ایم از نوعی از سنجش برای اندازه گیری یک بیومارکر مودر نظر استفاده کرده بودند، 9 تا گزارش مسئله داشتن با شناسایی محدوده را گزارش کرده اند. درعمل، سنجش ها محدوده های پایین تر شناسایی را داشته و داده های نتیجه شده از اینرو به سمت چپ سانسور

شده است. روشهای خاص مدیریت داده های سانسور شده به سمت چپ باید در این موقعیت ها درنظر گرفته شود و شامل هر دو رهیافتهای مبتنی بر احتمال و مبتنی بر MI می شود. چون داده ها به احتمال بیشتری برای مقادیر حداکثر و حداقل سانسور می شوند (خارج از طیف سنجش)، داده ها NMAR می باشند. رهیافتهای متداول رفتار ، داده ها را به صورت داده های از دست رفته یا نسبت دادن یک به یک به شکل 0 (همانگونه که 9 مطالعه در ارزیابی ما چنین کردند)، محدوده شناسایی یا نیمی از محدوده شناسایی تلقی کرده است. این رهیافت ها مشخص شده که منجر به تخمین های سوگیرانه شده که در آن سوگیری حین اینکه نسبت مشاهدات سانسور شده افزایش می یابد، افزایش یافته است. کار قابل ملاحظه ای به این حیطه از مطالعه اختصاص داده شده است (برای مثال رفرانسهای 26تا28) . برای روشهای خاص داده های سانسور شده به دلیل محدودیت های شناسایی، ما خواننده را به کار Hughes در رفرانس 27 ارجاع می دهیم که در آن عملکردها میآن روشهای مبتنی بر نسبت دادن و مبتنی بر احتمالات مقایسه شده است، یک روش مبتنی بر احتمال توصیه شده است و نرم افزار می تواند با ای میل زدن به نویسنده کسب شود.

خط مشی های عملی مدیریت داده های از دست رفته

در ذیل ما یک سری مراحل را برای ترکیب با یک آنالیز هنگام مواجه با داده های از دست رفته ارائه می دهیم.

مرحله 1: جمعیت مطالعه مورد هدف را شرح دهید: توصیف افراد واجدالشرایط برای مطالعه باید ارائه شود وگرنه مرتبط برای تعریف آنها باشد. (برای مثال افرادی که سرطآن پروستات تایید شده بافت شناسی دارند) نباید شامل معیارهای مرتبط به موجودیت نتایج یا داده های متغیر همزمآن باشد.

مرحله دوم: به وضوح انحراف از مجموعه داده های تحلیلی را شرح دهید. مطالعات اپیدمیولوژی مولکولی اغلب از منابع داده های موجود کسب می شود و در این خصوص، منابع اصلی باید یا شرح داده شده یا ارجاع داده شوند. تفسیر قطعی تا معتبر نتایج ولیکن انحراف از نمونه تحلیلی است یعنی اینکه چه کسی وارد و چه کسی خارج می شود. به طور ایده آل، این امر باید با شرح جمعیت مطالعه منطبق باشد. Jordan و همکارانش در مطالعه خود که مرتبط با تماس استروژن با دوزبالا طی نوجوانی به دانسیته

ماموگرافیکی در بزرگسالی بوده است، یک نمونه عالی از شرح روشنی را در یک تصویر فلوچارتی ارائه کرده اند مبنی براینکه کدام بیمارآن واجد شرایط مطالعه اند و کدامیک از بیمارآن نهایتا در آنالیز داده ها وارد می شوند .

مرحله سوم: مشخصات جمعیت افراد مورد مطالعه را در مجموعه داده های تحلیلی از جمله داده های از دست رفته و یا سانسور شده شرح دهید. اغلب شرح جمعیت مطالعه ارائه می شود. ولی در مواردی که در آن تخمین های نقطه ای برای شرح روابط تنها طبق موارد کامل مشتق شده است (زیرمجموعه جمعیت مطالعه)، شرح مجموعه داده های تحلیلی برای تفسیر حیاتی است به خصوص اگر از دست رفتن داده ها سیستماتیک باشد.

مرحله چهارم: تفاوت های میآن مطالعات با و بدون داده های متغیرهای کلیدی را از لحاظ مشخصات جمعیت شرح دهید. علاوه بر کمک کردن در تفسیر نتایج، نیز به شکل ارزیابی فرضیه MCAR عمل می کند. با استخراج از مثال قبلی تر ، Jordan و همکارانش مشاهده کردند که شرکت نکردگآن واجدشرایط مطالعه (یا آن دسته داده های از دست رفته درمورد دانسیته ماموگرافی) از شرکت کنندگآن در سن مصاحبه متفاوت بودند اما نه از لحاظ قد، اندیس توده بدنی یا BMI، و سابقه داشتن یک بیوپسی سینه.

مرحله 5: فرضیات احتمالی برای مکانیسم داده های از دست رفته را مطالعه کنید.

- فرض کنید که داده ها از نوع MCAR باشد اگر هیچ شواهدی از تخطی آن وجود نداشته باشد (همانگونه که در مرحله 4 تعیین شده است) و هیچ مکانیسم معینی برای تولید داده های NMAR وجود ندارد.
- فرض کنید که داده ها از نوع MAR باشد اگر از شرایط MCAR تخطی صورت گرفته باشد، هیچ مکانیسم معینی برای تولید داده های NMAR وجود ندارد و متغیرهای کاندیدای کمکی وجود دارد.

- فرض کنید که داده ها از نوع NMAR است اگر یک دانش قبلی نشآن دهد که مقادیر از دست رفته مرتبط با مقادیر مشاهده نشده باشد. وضعیت های متداولی که برای داده های NMAR رخ می دهد زمانی است که افراد مورد مطالعه قادر به فراهم کردن سنجش ها به دلیل مقادیر غیرمشاهده شده نباشند یا نخواهند چنین کنند (برای مثال برای ارزیابی خیلی مریض باشند و شدت بیماری اندازه گیری شود) یا ذخیره اندازه گیری متغیر نتواند مقادیر خارج از طیف آن را ارزیابی کند (برای مثال سنجش های با محدودیت های شناسایی) .

دو مورد اخیر نیاز به کیفیت سنجی بیشتری دارد. برای اولین مورد، اگر شرایط MAR به نظر از لحاظ تئوریکی عملی باشد، فرد باید حضور متغیرهای کمکی را درنظر بگیرد که این فرضیه را احتمالا در عمل خواهند داشت. برای مثال تحلیل انجام شده توسط Jordan و همکارانش نشآن می دهد که داده ها MCAR نیست. اینکه آیا داده ها MAR باشد یا اینکه آیا امکآن داشته باشد افرادی که دانسیته سینه خاص دارند بیش و کم مستعد به شرکت در مطالعه باشند، قابل تعیین نیست. اما احتمالا به نظر نمی رسد که دانسیته سینه مرتبط با شرکت باشد و از اینرو منطقی است فرض کنیم که داده ها MAR باشد. ولیکن داده های کمکی هنوز برای برقراری این شرط لازم است . داده هایی که با دانسیته سینه همبستگی دارند (برای مثال BMI) و یا از دست دادن داده ها یک نقش مهمی را در براوردن این شرط ایفا می کند و نیاز به ترکیب دارد. برای دومی، حتی اگر شرط NMAR مورد ظن باشد، متغیرهای کمکی می توانند باز به شرط MAR امکآن برقراری دهند. برای مثال اگر شدت بیماری طی چندین نقطه زمانی اندازه گیری شود، و یک دانش قبلی نشآن دهد که انهایی که مقادیر از دست رفته دارند به احتمال بیشتری افرادی هستند که انقدر مریضند که به درمانگاه برای ارزیابی نمی ایند، انگاه شرط NMAR باید مورد سوظن قرار گیرد. اما ارزیابی ها در ویزیت های دیگر ممکن است با نمرات از دست رفته در نقطه زمانی خاص همبستگی یابد و می تواند به شکل متغیرهای کمکی بکار رود و شرط MAR بعد از مشروط سازی روی این متغیرها عملی تر می شود. از سوی دیگر، شرط MAR ممکن است در مواردی مناسب نباشد که برای آن طراحی مطالعه نیاز به تنها یک ارزیابی شدت بیماری به ازای هر بیمار دارد.

مرحله 6: یک آنالیز CC اجرا کنید. اگر فردی مفروض دارد که داده ها MCAR است و یک نسبت اندکی از داده ها از دست رفته باشد، این کار می تواند به شکل تحلیل اولیه و انحصاری عمل کند. در غیراینصورت انالیزهای اضافی باید انجام گیرد(مرحله هفت)

مرحله هفت: یک رهیافت تحلیلی دیگر درصورت نیاز انتخاب کنید. یک تحلیل استاندارد MI یک انتخاب منطقی تحت شروط MAR است (یا MCAR در حضور یک نسبت بزرگ از داده های از دست رفته برای بازده های احتمالی در کارایی). اگر NMAR محتمل باشد (حتی پس از درنظرگیری متغیرهای کمکی)، کاربردهای خاص روشهای مبتنی بر MI می تواند باز مورد استفاده قرار گیرد، اما مکانیسم داده های از دست رفته باید مدلسازی گردد و چندین مدلهای NMAR باید اعمال گردد. این مدلها در هر دو اجرا و قابلیت دسترسی نرم افزاری پیچیده تر است و به تفصیل بیشتری توسط سایرین مورد بحث قرار می گیرد. اگر دست کم یکی از متغیرهای از دست رفته همانی باشدکه منوط به سانسور به دلیل محدودیتهای شناسایی است، یک رهیافت مبتنی بر احتمالات بنا به شرح Lyles و همکارانش و Hughes توصیه می شود.

مرحله هشت: روشهای داده های از دست رفته دیگر را اجرا کنید. متغیرهای کمکی باید برای عملکرد بهینه درنظر گرفته شوند. به طور اختصاصی تر، فرد نیاز دارد مطمئن شود که متغیرهای مشخص شده در مدل نسبت دادن شامل 1) کلیه متغیرها در مدلهای علمی از جمله متغیر وابسته است 2) هر گونه متغیر مرتبط با از دست دادن داده های متغیرهاست و 3) هر گونه متغیرهایی که با متغیرهایی همبسته است که داده های آن از دست رفته است از اینرو MAR می تواند درست باشد. هنگام مواجه با انتخاب اینکه کدام متغیرهای کمکی در نسبت دادن متغیرهای مورد نظر گنجانده می شود، Collins و همکارانش از طریق مطالعات شبیه سازی نشآن دادند که به حساب امدن بیشتر حتی زمانی که از فایده برخی متغیرهای کمکی مطمئن نیستیم منجر به کارایی افزایش یافته و کاهش سوگیری ها می شود. در مطالعه توسط Jordan و همکارانش، این کار دربرگیرنده کلیه متغیرها در مدل علمی (که با استفاده از رگرسیون پلکانی به طرف جلو انتخاب شده است)، نتیجه (سنجش ماموگرافی)، سن در هنگام مصاحبه می باشد چون

با از دست دادن داده ها و هر گونه متغیرهای همبسته با سنجش ماموگرافی (برای مثال BMI) همبستگی دارد. هرچند قد و تاریخچه بیوپسی پستآن مرتبط با از دست دادن داده ها نیست، آنها باید گنجانده شوند اگر مرتبط با سنجش ماموگرافی بوده یا اینکه در مدل علمی گنجانده شده بوده اند.

شکل خاص مدل با انتخاب در نرم افزار مشخص خواهد شد که در بخش محدودیت ها مورد بحث قرار گرفت. همانگونه که van Buuren گفته است، این انتخاب بستگی به این امر دارد که آیا متغیرهای مورد نظر از نوع مختلط می باشند (یعنی برخی ترکیبات متغیرهای پیوسته، طبقه بندی شده، رتبه ای، یا دوتایی) یا اینکه یک توزیع مشترک واقع گرایانه را می توآن مشخصه سازی نمود. اگر دومی درنظر گرفته شود، نرم افزاری که باید از رهیافت مدلسازی مشترک استفاده کند (برای مثال SAS) توصیه می شود. درغیراینصورت نرم افزاری که رهیافت مشخصه سازی شرطی کامل را بکار گرفته است توصیه می شود (برای مثال STATA یا R).

مرحله نه: یک تحلیل حساسیت را اجرا کنید. اگر روشهای داده های از دست رفته بکار گرفته شود، فرد باید یک تحلیل حساسیت را با متناسب سازی مدلهای مختلف بکار گیرد. اگر نسبت از دست رفتن داده ها کوچک نباشد و داده ها MCAR باشد، هر دو تحلیل های CC و MI را می توآن انجام داد که در آن شرایط MAR مختلف (یعنی مجموعه های مختلف متغیرهای کمکی) درنظر گرفته شوند. اگر NMAR مورد تردید است، رهیافت های CC، MI با استفاده از مجموعه های مختلف متغیرهای کمکی تحت MAR، و کاربرد خاص رهیافتهای MI تحت چندین مدل NMAR باید اجرا گردد. ارایه نتایج را می توآن برای شرح استحکام یافته ها در میآن فرضیات متعدد ارائه داد. بعلاوه، عدم یقین این مفروضات می تواند زمانی توجیه شود که یافته های خلاصه سازی شده ارائه گردد. برای نکات و مسائل دیگر درمورد اجرای MI از جمله اینکه چند نسبت دادن انجام بگیرد به کتاب خوب Allison درمورد داده های از دست رفته رجوع کنید که خط مشی های عملی مفصلتر و مثالهایی را برای بکارگیری MI به زبآن غیرفنی ارائه می دهد.

مرحله ده: نتایج را تفسیر کنید. اگر یک اتفاق نظر به سهولت در میآن انالیزها کسب شود، این امر تفسیر (و ارائه) را مستقیم می سازد. در موقعیت هایی که در آن یافته ها بنا به مفروضات متفاوت است، نویسندگآن

مجبورند درباره اینکه چه کاری عملی ترین است بنا به فرض درکشآن از بیولوژی سنجش انجام دهند و نتایجی را ارائه دهند که خواننده بتواند درک کند چگونه مفروضات بر کل تفسیر اثر می گذارد.

نتیجه گیری ها

خلاصه اینکه، ما نشآن داده ایم که مطالعات اپیدمیولوژی مولکولی با یک مسئله داده های از دست رفته بویژه چالش برانگیز در این خصوص روبرو است که اکثریت این مطالعات داده های از دست رفته درمورد متغیر کلیدی مورد نظر یعنی بیومارکر ها را دارند.

هرچند به نظر معنی دار می رسد که تنها روی افرادی مطالعه کنیم که بیومارکرهای اندازه گیری شده دارند، ما روی اهمیت گنجاندن افرادی بحث می کنیم که برای مطالعه واجدشرایط هستند علی رغم اینکه بیومارکر از دست رفته دارند. در حداقل موارد، ما روی مقایسه خصوصیات بین افرادی بحث می کنیم که داده های از دست رفته و از دست نرفته دارند. ما قویا ترکیب روشهای داده های از دست رفته را با تحلیل هنگامی که تضمین شده باشد، تشویق می کنیم. به طور اخص تر، اگر مقایسات نشآن دهد که داده ها MCAR نیستند، و MAR به نظر منطقی اید، ما قویا استفاده از MI استاندارد را توصیه می کنیم. حتی در مواردی که در آن داده ها MCAR باشند، فرد می تواند از MI از لحاظ کارایی نفع برد. اگر محتمل باشد که داده ها NMAR باشند و فرد بتواند درصد اطلاعات کمکی قوی را مفروض دارد، MI استاندارد ممکن است باز یک ابزار تقویت کننده تخمین منطقی باشد. در غیراینصورت MI که مکانیسم داده های از دست رفته را مدلسازی می کند یک احتمال است. یک خصوصیت مفید MI آن است که به ترکیب عدم قطعیت این فاکتورها به شکل نتایج امکآن می دهد.

پیوست A: برنامه ریزی کدبندی شده STATA برای اجرای MI

```
/*Read in data set where interest is in estimating effects
  of 2 risk factors on case-control status*/

/* X1 and X2 are risk factors of interest and Z is a potential
   auxiliary variable */

    insheet using "~/scen1.csv",
    clear

/*Use ICE to fit imputation model creating 10 imputed
  data sets*/
/* All variables in scientific model are included in impu-
   tation model in addition to auxiliary variable */

    ice case x1 x2 z, saving(simimpute.dta) m(10) replace

/*Read in data set containing all 10 imputed data sets*/

    use simimpute.dta, clear

/*Use MICOMBINE to fit the desired scientific model (a
  logistic regression model that includes the risk factors
  of interest) and combine results across 10 data sets*/

    micombine logit case x1 x2
```