האם רוב התגליות המדעיות שגויות?

מאת **פרופ' יואב בנימיני**

הכותרת למאמר זה אינה מקורית. זוהי כמעט כותרתו של מאמר מאת הרופא והאפידימיולוג איונידיס (Ioannidis)[1]. הכותרת המקורית של מאמרו הייתה דרמטית אף יותר: "מדוע רוב התגליות המדעיות המתפרסמות שגויות?" במאמר עצמו טען להוכחה מתמטית שאכן כך הדבר ונתן הסברים ארוכים לתופעה. אף שלא היה הראשון לטעון זאת, מהר מאוד הכה מאמר זה גלים מחוץ לקהילה המדעית. כשנתיים לאחר פרסומו הוא צוטט בספרות המדעית רק כמה עשרות פעמים (כיום אלפים), אבל הורד כ־100,000 פעם ונצפה יותר מחצי מיליון פעם. ה"בוסטון גלוב" ציין את העיסוק האינטנסיבי והאובססיבי במאמר כלידתה של כת תרבותית, ואכן כך היה.

ה"ניו יורקר" שאל בכתבה גדולה בשנת 2010: **Is there something wrong with the scientific method?** וכותרות נוספות בעיתונות הלא מדעית היו "Unreliable Research", "Trouble at the lab", והיו גם נחרצות יותר: "How science goes wrong".

טענתו של איונידיס עלתה לאחר שבתחומי מחקר שונים נכשלו מחקרים חדשים בשחזור תוצאותיהם של ניסויים קודמים. כך היה במחקרים רפואיים ואפידמיולוגיים, בחקר המוח, בחקר ההתנהגות ובמחקרים גנטיים שבהם מנסים לקשור בין סמנים על גבי הגנום לבין מחלות כמו סוכרת או יתר לחץ דם על מנת להבין את המנגנונים המעורבים במחלה.

גם במדע השימושי יותר קיימים סימנים לבעיה: לפני כמה שנים התריעו ברשות האמריקאית לתרופות ומזון (FDA), המופקדת על רישוי תרופות חדשות, שכמחצית התרופות החדשות שהגיעו לשלב הניסויי השלישי והמכריע לא עברו אותו בהצלחה ולא הורשו לשימוש. כדי להבין את חומרת התופעה צריך לדעת שניסוי מגיע לשלב השלישי אחרי סדרה לא קצרה של ניסויים קודמים שפורסמו בספרות המדעית. סכומי הכסף המושקעים בניסויי השלב השלישי אדירים, שכן הניסוי מנוהל בקפדנות רבה ומערב חולים רבים. יתר על כן, אם אין תועלת בתרופה המוצעת, הרי שלשווא נחשפו מאות ואלפי חולים לטיפול הניסיוני, ואם התרופה מיטיבה עם החולה אך נכשלת ברישוי, התוצאה בוודאי גרועה.

התייחסות הממסד המדעי לבעיית הממצאים שאינם ניתנים לשחזור כאל בעיה מערכתית ולא נקודתית במחקר מסוים הגיעה מאוחר יותר, אולם בארבע השנים האחרונות מעסיקה רבים בתוך העולם המדעי. עיתוני המדע הרב־תחומיים *Nature* ו־*Science* התייחסו לנושא במאמרי מערכת, וכך עשו גם עיתונים מובילים אחרים. גופי מחקר אמריקאיים כמו המכון הלאומי לבריאות (NIH) והקרן הלאומית למדע (NAS) הקימו צוותי חשיבה בנושא. בפברואר 2017 תקיים הקרן סדנה, זו הפעם השנייה, בניסיון נוסף להציף את הבעיות שבבסיס התופעה ולדון בפתרונות אפשריים. גם בארץ נערכה בינואר 2015 סדנה בנושא זה בנוגע להדירותם (replicability) של ניסויים הנערכים בבעלי חיים.

מהן הסיבות לתופעה? בהתבטאויות של מי שאינם מדענים, אך גם של מדענים רבים, מיוחסת הבעיה לגורמים הקשורים לסוציולוגיה של המדע ולפסיכולוגיה של מדענים. למשל: חשיבותם של הפרסומים המדעיים לקבלת קביעות במוסדות אקדמיים ולהתקדמות בדרג האקדמי או הצורך לפרסם תוצאות מרשימות על מנת לזכות במענקי מחקר מעודדים חוקרים לפרסם תוצאות שאינן מבוססות דיין. לעתים נדירות התופעה מגיעה עד למעשה מרמה של ממש, ומקרים כאלה זוכים לתגובה זועמת ונחרצת, למשל Stapel מטילבורג שבהולנד ו־Hauser מהרווארד אולצו לפרוש ממשרותיהם לאחר שהתגלה שינוי נתונים במחקריהם. עם זאת החשש מהשפעת פרשנות יתר, מהדגשת התוצאות התומכות במסקנות המחקר בלבד ואפילו מהטיה לא מודעת בפרסום – קיים תמיד.

סיבה נוספת היא העניין של עורכי העיתונות המדעית בממצאים חיוביים מרעישים, למשל פרסום מחקר המוצא שאכילת פיצה מקטינה את הסיכוי לסרטן הערמונית מעניין יותר מדיווח על שלא נמצא קשר בין אכילת אי אילו מאות מאכלים לבין שיעור התחלואה בסרטן זה. ואכן, בטובים שבעיתונים המדעיים מטילים איפול על תוכנם של מאמרים שהתקבלו לפרסום עד יום הפרסום עצמו, ובינתיים מכינים את מערכת יחסי הציבור עבורם. תחרות בקרב המובילים שבעיתונים המדעיים – מי יפרסם מאמר מרעיש עולמות (ורב־ציטוטים) – דומה לזו הקיימת בעיתונות הרגילה.

יש המאשימים את האינטרסים הכספיים של מדענים בתוצאות מחקריהם. יש הטוענים רק נגד נטייתם של חוקרים לרצות גופים המממנים מחקריהם, מבלי להבין שבכך נוצרת הטיה, למשל במתן אישור לחברת תרופות לגנוז פרסום תוצאות מאכזבות. לעתים האינטרסים הכספיים המואשמים הם של האוניברסיטאות, המנסות להתגבר על תקציביהן המידלדלים באמצעות חברות היישום הנשענות על תוצאות החוקרים.

נראה כי הדוגמה הבאה מאששת טענות אלו. חוקרים מאוניברסיטת Duke פיתחו אלגוריתם המנצל נתונים של התבטאות גנים של חולים בסרטן הדם לבחירת משטר הטיפול המתאים. תוצאות טיפול זה המותאם־אישית, שהתפרסמו ב־2006, נראו מרשימות, והאוניברסיטה הקימה חברת יישום שתמסחר את גישת הטיפול הזאת ואף החלה בניסוי קליני לבדיקתה. אלא ששני סטטיסטיקאים ממרכז המחקר אנדרסון לחקר הסרטן ניסו לשחזר את השיטה מתוך המידע שבמאמר ונכשלו. הם גילו בעיות של אי־תאימות בין התוצאות המדווחות לבין הנתונים וניסו להתריע על הבעיות בעיתונים המרכזיים שבהם התפרסמו המחקר המקורי ומחקרי ההמשך, אך ללא הצלחה. לבסוף פרסמו ביקורתם בעיתון סטטיסטי.[2] הביקורת זרעה ספקות בקרב המשקיעים, אולם רק לאחר שהתברר שהחוקר הראשי התגאה בעיטור כבוד שבו לא זכה, נערכה בדיקה מעמיקה שבעקבותיה נאלצו החוקרים לסגת ממחקרם, והאוניברסיטה – מיזמתה. ◄

◄ מקרים כגון אלה זוכים לפרסום נרחב ומחשידים ציבור רחב של מדענים, וחמור מזה, מעוררים ביקורת על המדע בכללותו, כפי שנאמר בפתח דבריי. אלא שתופעות דרמטיות שכאלה הן מועטות. יתר על כן, כל הנימוקים הסוציולוגיים הללו מתעדים תופעות הקיימות זה זמן רב ושלא השתנו בעשור האחרון. פרסום, כבוד וכסף תמיד היו כוח מניע של מדענים. אם כך, אם הבעיה החמירה, ואמנם קיימת עלייה במספר התגליות המדעיות השגויות, ההסבר לבעיה זו, ולכן גם הטיפול בה, צריכים להיות שונים.

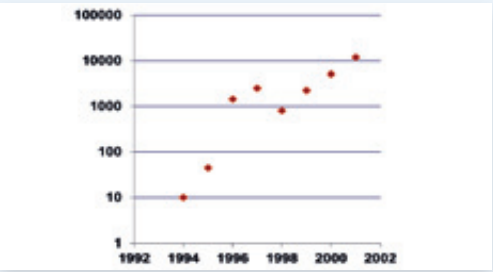
ואכן, לדעתי, ההסבר הוא שבתחומים רבים העשייה המדעית עוברת שינוי דרסטי. בתיאור כוללני ופשטני במקצת הייתי אומר שהמחקר המדעי עובר תהליך של תיעוש. תהליך התיעוש זוכה לשמות שונים: בתחומים הנשענים על נתונים קיימים, כמו אפידמיולוגיה, כלכלה, סוציולוגיה ותקשורת, השימוש בכמויות נתונים אדירות זוכה לכותרת Big Data – נתוני עתֵק – כמו גם Data Mining – כריית מידע. במדעים הניסויים נשענים על שיטות High Throughput, שבהן הניסויים וניתוח תוצאותיהם נעשים באוטומציה כמעט מלאה. כאשר מכשור יקר, ייחודי ועתיר תוצאות משמש למחקרים בפיזיקה או באסטרונומיה, הליך המחקר זוכה לכותרת Big Science (מדע גדול).

אם נשווה את השינויים הללו לאלה שעברה תעשיית המכוניות, ניתן להבין מדוע אני מכנה זאת תיעוש העשייה המדעית. ייצור המכוניות בעלות מנוע השרֵפה הפנימית החל ב־1888 אצל יצרנית המכוניות בנץ (לימים: מרצדס–בנץ) יוצרו חמש מכוניות בשנה, בעבודה יחידנית. ב־1902 יושם לראשונה פס ייצור סדרתי של מכוניות אצל אולדסמוביל. מן הפס הזה ירדה מכונית כל שעתיים, ובסך הכול כ־1,500 בשנה. פורד שיפר בהרבה את תהליך הייצור הסדרתי, וב־1914 יוצרו ממודל T המפורסם כארבע מכוניות בשעה, ובסך הכול כ־12,000 מכוניות בשנה. פס הייצור הרובוטי החל ביפן בשנות השמונים של המאה הקודמת ושוכלל מאז לעין ערוך. כיום זו הדרך שבה מיוצרים שבעים מיליון מכוניות חדשות מדי שנה, בהתערבות קטנה ביותר של מפעילים אנושיים השולטים על התכנון ועל התהליכים, אך לא על הביצוע.

במדע מתרחש תהליך דומה. ניקח למשל מחקרים הבודקים אם התבטאותם של גנים היא שונה ממצב למצב, למשל בתא סרטני לעומת תא שפיר. בדגימת מאמרים שערכנו בעיתון מוביל מתחום הגנטיקה נמצא כי במאמר שהתפרסם ב־1994 דיווחנו על התבטאותם של כ־4–10 גנים. ניתוח ההתבטאות של כל אחד מהגנים היה תהליך ארוך וממושך של עבודה ידנית ויחידנית. ב־1995 מצאנו במאמר התבטאות של כ־80 גנים. ב־1996 דווח במאמר על כ־1,000 התבטאויות



תיעוש תהליך ייצור המכוניות - מימין: 1888 מנוע הבערה הפנימית של בנץ - 5 מכוניות בשנה בייצור ידני, 1914 מודל T של פורד - 12,000 מכוניות בשנה מפס ייצור, פס ייצור רובוטי - 70 מיליון מכוניות בשנה



מספר הגנים שהתבטאותם מדווחת (על הציר המאונך) במאמר שנדגם בשנה מסוימת

גנים. מספר זה של התבטאויות הושג הודות לעובדה שבתקופה זו החלו משתמשים בטכנולוגיות ניתוח חדשות (מיקרואריי) שבהן ה"ניסויים" נעשים אוטומטית על מספר רב של גנים או סמנים בעת ובעונה אחת ובקו ייצור סדרתי. התוצאה היא המשך גידול מדהים: ב־1997 כבר נמצאו 4,000 התבטאויות גנים, ב־2001 – 12,000, וכיום מדידה בו־זמנית של התבטאות כל הגנים, כ־20 אלף הגנים באדם, הוא הסטנדרט כמעט בכל ניסוי. כשמדובר על סמנים גנטיים על הגנום כולו, הניתוח הרגיל בימים אלו כולל כחצי מיליון סמנים ומגיע לשני מיליון. במחקר שעשינו לאחרונה חיפשנו קשרים מעניינים בין סמנים גנטיים לגודלם הפיזי של אזורים במוח. החיפוש היה על פני 13.5 ביליון קשרים אפשריים.

ואם בתעשיית המכוניות התרחשה המהפכה במשך 100 שנה, במחקר המדעי היא התרחשה בפחות משני עשורים (!)

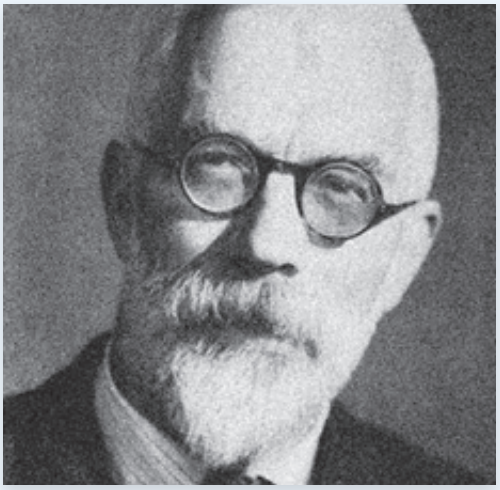
נחזור לייצור המכוניות – מה בדבר איכותן? האם היום, כאשר ייצורן נמדד במיליונים, משתמשים באותן שיטות לאבטחת איכות ששימשו בעבר, כאשר הייצור היה של מכוניות בודדות? ברור שלא. אצל בנץ ליטף האומן את ירכיה של כל מכונית חדשה, נסע בה נסיעת מבחן והריח את שמן המנוע כדי לוודא שאין בה פגם. שיטות לבקרה תהליכית סטטיסטית על מנת לשמור על איכות בפסי הייצור נכנסו לשימוש רק בתקופת מלחמת העולם השנייה. תעשיית המכוניות היפנית הבינה שאין די בכך, ובשנות השמונים החלה להשתמש בשיטות של ניהול לאיכות כוללת – TQM (Total Quality Management). ואכן איכותן המיוחדת של המכוניות היפניות בשנים ההן גרמה שגישות TQM מצאו את דרכן לכל העולם, וכיום הן מוטמעות בכל תהליכי הייצור והשירות המודרניים.

אשר למדע, בבסיס השיטה המדעית קיימת ההכרה שעל מסקנות להיות מבוססות על נתונים אמפיריים. עם זאת הנתונים הם תמיד חלקיים ולא מדויקים, ולעולם תיתכן שגיאה.

כפי שמדענים למדו להסתמך על קירוב באמצעות מודלים מתמטיים לתיאור המציאות, כך למדו הם להסתמך גם על הסטטיסטיקה על מנת לתחום את השגיאה בהעלאת מסקנות מניסוי אמפירי. ◄

תיעוש התהליך המדעי

◄ לפני כמאה שנה ביסס הסטטיסטיקאי והביולוג האבולוציוני פישר את העיקרון שעל מנת שתוצאה של ניסוי תיחשב מבוססת, עליה להיות מובהקת סטטיסטית עם שיעור שגיאה של פחות מ־5%. כלומר, גם אם אין כל תופעה ממשית מאחורי התוצאה, והיא נולדה באקראי ומשגיאות מדידה בלבד, קיים סיכוי, אמנם קטן, שהתוצאה תעבור את הרף שהציב פישר. עיקרון זה שימש את המדע היטב עד לאחרונה.



סר פישר: "שמור על הסתברות קטנה לקבל תגלית שגויה בעת בדיקת ממצא אמפירי בודד".

אבל אם בודקים עשרות, מאות ואף אלפי תוצאות אפשריות ובוחרים מהן את התגליות על פי העיקרון שמגביל את הטעות ל־5% לכל תוצאה לחוד, מספר השגיאות הצפוי הולך וגדל. בסריקת אלף תוצאות מסתתרות אולי עשרים תגליות מדעיות אמתיות, וכל היתר יכולות להוביל רק לתגליות שווא. לכן נצפה למצוא 49=5/100*(1000–20) תגליות שיעברו את רף המובהקות של 5% למרות היותן שגויות. גם אם נצליח לכלול בתגליותינו את כל 20 התגליות הנכונות (ובדרך כלל איננו מגיעים לכך), עדיין רוב התגליות יהיו שגויות: 49 מתוך 69. אם לא נתאים את השיטות הסטטיסטיות למצב החדש, מספרן של התגליות השגויות יאפיל על מספרן הקטן של התגליות האמתיות שאותן אנו מחפשים, ושיעור התגליות השגויות יהיה גבוה.

הטרוניה של איונידיס (וכותרת המאמר) מבוססת בדיוק על בעיה זו. הוא טוען למשל שככל שתחום מחקרי "חם" יותר, כן התגליות השגויות שבו רבות יותר. שטח חם פירושו הוא זה: חוקרים רבים יותר, וכשהחוקרים רבים יותר, המחקרים רבים יותר, וכשהמחקרים רבים יותר, אפשרִיות תגליות רבות יותר, ואתן שגיאות רבות יותר. מבלי משים איונידיס מסביר כיצד השיטות הישנות נכשלות בבואן להתמודד עם בעיות גדולות. התוצאה המתסכלת היא האשמה המוטלת על כתפי הסטטיסטיקה כשלעצמה ולא על השימוש הלא נאות בשיטותיה: כך למשל עיתון בתחום הסוציולוגיה (אמנם לא מרכזי) הטיל חרם על השימוש בגישת המובהקות הסטטיסטית ודרש מהכותבים בו להימנע מכל שימוש בשיטות סטטיסטיות לבד מתיאור הנתונים. הלחץ הביא את האיגוד האמריקאי לסטטיסטיקה לפרסם באופן חריג נייר עמדה באשר לדרך השימוש במובהקות סטטיסטית. לצערי הנייר עסק רק במובהקות סטטיסטית והצביע על הזהירות הדרושה בעת השימוש בה, ורק רמז על הבעיה העיקרית המשותפת לכל השיטות הסטטיסטיות המסורתיות בעידן המדע התעשיתי: הכמות העצומה של הממצאים הנבחנים, שמהם מסוננים, מדווחים, ומודגשים כתגליות רק מעטים.

ראוי לציין שכבר בשנות החמישים של המאה הקודמת התעורר צורך ראשוני להתמודד עם ניסוי שבו בוחנים יותר מתוצאה אפשרית אחת, כאשר משווים בין יותר משני טיפולים אפשריים. שיטות שפיתחו טוקי (Tukey), שֶׁפֶּה (Scheffe) ואחרים, ובארץ חוקרים כמו פרץ, מרכוס, גבריאל והוכברג, הפכו למקובלות בתעשיית התרופות, שבה השוואות בין מספר קטן של טיפולים הן דבר שבשגרה. אלא ששיטות אלו

דורשות הורדה דרסטית של סף מובהקות הנדרש להוכחה כדי לשמור את ההסתברות לאפשר אפילו תגלית שגויה אחת מתחת ל־5%. כשניסו להתמודד עם בעיות גדולות יותר, גרמו דרישותיהן המחמירות פגיעה בעצמת המחקרים והם בדרך כלל נזנחו.

פרופ' טוקי: "שמור על הסתברות קטנה לקבל ולו גם תגלית שגויה אחת בעת בדיקת מספר ממצאים אמפיריים יחדיו".

ובכל זאת גישות סטטיסטיות חדשות המתאימות לעידן המדע התעשיתי קיימות. ניתן לקחת את מאפיין הביקורת שבו השתמשו איונידיס וקודמיו ולהפכו לקריטריון מטרה. לשם כך נגדיר כקריטריון את שיעור התגליות השגויות מבין התגליות שנעשו במחקר, ופורמלית – את תוחלת היחס שבין מספר התגליות השגויות למספר התגליות (False Discovery Rate - FDR). לפי הקריטריון החדש, תגלית שגויה אחת, ואפילו שתיים, במחקר שבו נמצאו ארבעים תגליות הוא מצב קביל, אבל כשמדובר בתגלית שגויה אחת מתוך ארבע תגליות זהו שיעור גבוה מדי, ומצב זה אינו קביל. אם כן, במקום לא לאפשר ולו תגלית שגויה אחת, ניתן להסתפק בשליטה על שיעור התגליות השגויות. ניתן כעת לפתח שיטות השולטות ישירות על שיעור התגליות השגויות הרצוי בהצבת רף גבוה יותר לקביעת תגלית, ועם זאת לא להחמיר מדי ולא לפגוע שלא לצורך בעצמת המחקר.

קריטריון ה־FDR והשיטה הבסיסית ליישומו הופיעו ב־1995 במאמר שכתבתי עם יוסף הוכברג מאוניברסיטת תל אביב.[3] בשיטה זו מנמיכים בהדרגה את סף המובהקות הסטטיסטית לתוצאה בודדת – מ־p=0.05 (5%), המשמש לבדיקת המובהקות של תגלית בודדת, לרמות נמוכות יותר. לכל רמה p שכזו נחשב את מספר התגליות הצפויות לעבור בשגגה את הסף, שבדוגמה הוא 1000p (1000–20) או לכל היותר 1000p. כמו כן נספור את מספר התגליות שעוברות רף מובהקות זה, r(p). נחפש כעת את ה־p הגבוה ביותר, שבו

$$1000p\,/r(p) \le 0.05$$

משוואה 1

מפתיע אולי שאפשר לבסס רעיון אינטואיטיבי זה באמצעות הוכחה מדויקת, אך במאמר המקורי אכן הוצגה הוכחה, בהנחות מגבילות. בהמשך הושגו הוכחות גם בהסרת ההגבלות, על ידי קבוצות מחקר בתל אביב,[4] בסטנפורד[5] ואחרות.

האתגרים המעשיים הגדלים והולכים עם התפתחותן של שיטות מחקר מורכבות יותר הביאו לפיתוח שיטות שונות ומגוונות המנסות להבטיח שליטה על שיעור התגליות השגויות. מלבד זאת קבוצות מחקר רבות בסטטיסטיקה עוסקות כיום בבעיה הבסיסית שהצגתי: כיצד מתמודדים סטטיסטית עם אי־הוודאות הגדולה כאשר מאוסף האפשרויות הגדול בוחרים את התוצאות הספורות המבטיחות ביותר. משתמשים בשיטות אלו במידה רבה בבעיות הגדולות ביותר מקום שבו המדענים כבר מודעים היטב לצורך לשלוט על שיעור התגליות השגויות, ואכן המאמר המקורי[3] הוא במאה המאמרים המצוטטים ביותר בעולם המדעי. אולם קצב חדירתן של שיטות אלו עדיין אטי יחסית למהירות ההתפשטות של התהליך התעשייתי במדע. ◄

◄ ברפואה, באפידמיולוגיה, בפסיכולוגיה ניסויית, במדעי החברה ובמחקרים פרה־קליניים למשל השימוש בהן מזערי.

אם כך, האומנם יש בשיטה המדעית מכשלה, לפחות בתקופת המעבר בטרם אימוץ שיטות מתאימות יותר? לאו דווקא. עיקרון נוסף בבסיסה של השיטה המדעית מהווה את חומת ההגנה מפני מכשלות שכאלה, כמו גם מהמכשלות שנובעות מהסוציולוגיה של קהילת המדע שנזכרו קודם. תגלית מדעית נחשבת מבוססת היטב אם חוקרים נוספים על אלה שגילו אותה יכולים לחזור ולגלותה, במעבדה אחרת, באוכלוסייה אחרת, בבית חולים אחר, ואולי אפילו בשיטה קצת אחרת. זו למעשה הדרישה שהתגלית תהיה הדירה. ובנושא זה אין ויתורים: אין תגלית מדעית מקודשת. כל תגלית עומדת למבחן שוב ושוב, והספקנות הבסיסית הדרושה לכך היא תכונת יסוד חיונית למדען.

ההכרה בחשיבותה של בדיקת ההדירות שבה ועולה אף שבמדע גדול חזרה על ניסוי היא מבצע יקר שאינו מניב תוצאות הֶרואיות. בפסיכולוגיה ניסויית, תחום שאירעו בו כמה מהכישלונות המתוקשרים שהזכרתי, הסתיים לאחרונה מאמץ מרוכז של מדענים לחזור על כל הניסויים שתוצאותיהם פורסמו בשלושה עיתונים מובילים לפני חמש שנים. המחקר הוכיח שבעיית ההדירות היא אמתית: קרוב לשני שלישים מהמחקרים לא שוחזרו, כלומר לא עברו את סף המובהקות במחקר השחזור. במחקר טרום־קליני בסרטן נערכים כעת למבצע שחזור דומה של תוצאות 100 מאמרים (בהובלת ארגון Open Science Initiative). בחקר הרקע הגנומי למחלות כגון סוכרת, יתר לחץ דם וסכיזופרניה מוקמים מַאֲגָדים (קונסורציומים) גדולים הכוללים מספר רב של קבוצות מחקר, המנסים להעריך בעבודה משותפת אילו מהתוצאות מקבלות אישוש ביותר ממרכז מחקר אחד, ולכן עוברות את מבחן ההדירות.

אולם לא תמיד אפשרית חזרה על ניסוי לבדיקת ההדירות שתתבצע באופן בלתי תלוי בידי חוקרים נוספים, וגם כאשר בדיקת ההדירות אפשרית, הזמן העובר מזמן ההכרזה הראשונית על התגלית ועד לבדיקת הדירותה יכול להיות קריטי. קחו למשל את מחקרם של ריינהרט (Reinhart) ורוגוף (Rogoff), שני כלכלנים מהרוורד שפרסומם בא להם בין השאר משום שחזו את המפולת הגדולה בשוק הנדל"ן האמריקאי, מפולת שסחפה אחריה את הכלכלה העולמית. בשנת 2010 פרסמו שני הכלכלנים הללו עבודה אמפירית שממנה הסיקו שאם החוב הלאומי מגיע ל־90% מהתוצר הלאומי הגולמי, הצמיחה של המשק הופכת לנסיגה. הם קראו לאותה נקודה קריטית של היפוך המגמה Tipping point. בעקבות מסקנתם זו, כאשר כלכלת יוון התקרבה לנקודת מפנה זו, הטיל עליה הבנק האירופי אמצעי צנע חמורים כתנאי לעזרה. ההפגנות, בחלקן אלימות, זכו לסיקור נכבד וכך גם הפיטורים, האבטלה והשקיעה החברתית שנבעו ממנה. גם איטליה וספרד התקרבו לאותה נקודת אל־חזור מסתורית.

בינתיים כלכלנים אחרים ניסו לקבל תוצאה דומה באמצעים אחרים: נתונים אחרים, שיטות אחרות, ולא הצליחו – בעיית הדירותה של התוצאה. הניסיונות הכושלים האלו הביאו חוקרים צעירים מאוניברסיטת מסצ'וסטס לנסות לשים ידם על הנתונים ועל הניתוח המקורי של ריינהרט ורוגוף. לפני כשנתיים יצאו החוקרים בהכרזה שהתוצאה המקורית אינה ניתנת לשחזור. והסיבה? הניתוח המקורי נעשה באמצעות גיליון אקסל, ובעת שרוגוף וריינהרט בחרו את השורות שעליהן יריצו את המודל של רגרסיה, נשמטו כמה שורות מהבחירה. אם כוללים אותן שורות בניתוח ונמנעים מהנחות נוספות, שכעת ניתן להבחין בהן, נעלמת נקודת המפנה המאיימת: עדיין ככל שהחוב עולה הצמיחה יורדת, אך אין סיבה להתייחסות מיוחדת ל־90%. אכן, חוסר הדירות התוצאה חשף בעיה, אך העיכוב

של כמה שנים והסבל הרב שנגרם בינתיים היו יכולים להימנע לו היו החוקרים נוקטים פתיחות שהייתה מאפשרת לבדוק את שחזור מחקרם, מנתוניהם למסקנתם, מיד עם פרסום המחקר המקורי.

הלקח הוא שעל שקיפות בתהליך הניסוי ובניתוח הנתונים ועל הבטחת יכולתם של אחרים לשחזר את תהליך עיבוד הנתונים להיות הנדבך הראשון בהבטחת ההדירות (למען האמת, זו לא רק הבטחת יכולתם של אחרים, אלא ראשית – הבטחה שאותו חוקר יוכל לחזור ולקבל אותן תוצאות גם אחרי שנה). דרישה זו לאפשרות שחזור – כל הדרך העוברת מטופס איסוף הנתונים, או בסיס המידע, ועד לתרשים במאמר – נקראת reproducibility ומוכרת גם כדרישת שקיפות (Transparency). כמה עיתונים מדעיים חשובים דורשים אותה לאחרונה במידה רבה יותר, והיא מופיעה במרבית המלצות הוועדות השונות.

יזמת reproducible research, הפועלת בכיוון זה, משלבת יחדיו אנשי מדעי המחשב, מתמטיקאים שימושיים וסטטיסטיקאים במאמץ לפיתוח כלים שיאפשרו לתעד אוטומטית את מסלול החישוב וניתוח הנתונים.

לסיכום

שקיפותם של שיטות, נתונים ותהליכי ניתוח להבטחת יכולת השחזור של תוצאות הניסוי היא נדבך אחד להבטחת הדירות המחקר המדעי, והשימוש בשיטות סטטיסטיות המתאימות עצמן לאפשרויות הבחירה העצומות בהיקפן העומדות בפני החוקר במדע התעשייתי, שאותן סקרתי למעלה, הן הנדבך השני. הישענות על שני נדבכים אלו תבטיח שרוב התגליות המדעיות המדווחות לראשונה אכן יהיו נכונות ויעמדו במבחן ההדירות אם וכאשר יתבצע. ■

* מאמר זה מבוסס על מאמר קודם שהתפרסם בגיליון אפריל 2015 של "אודיסאה" (ז"ל) "תעשיית התגליות זקוקה לתיקון", שעליו נשענה גם הרצאתי באקדמיה הישראלית למדעים בחנוכה תשע"ו, ואני מודה לעורכי "אודיסאה" על ההזמנה ועל העזרה בעריכת המאמר המקורי.

* The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [PSARPS-294519]

1) Ioannidis JPA, Why most research findings are false. *PLOS Medicine*, 2,8, 696–701 (2005).

2) Baggerly K. and Coombes K., "Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology" Annals of Applied Statistics (2009).

3) Benjamini Y. and Hochberg, Y., Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57, 1, 289–300. (1995) (59th in citations all across Science, according to Nature 2014).

4) Benjamini Y., Heller, R. and Yekutieli, Y., Selective Inference in Complex Research. *Philosophical Transactions of the Royal Society A*,367, 4255–4271 (2009).

5) Efron, B., "Large Scale Inference", *Cambridge University Press*, Cambridge (2010).