

مقالة رأي وتحليل

التاريخ التحريري: ١ نوفمبر ٢٠٢٦

التنبؤ الذي غيّر المستقبل الذي تنبأ به

> كان الإنذار يُخبرنا بما قد يحدث. ثم قررنا أن نتصرف بناءً عليه. وحين لم يحدث ما حذّر منه، طلبنا منه أن يدافع عن صدقه. كأننا نحاكم رجل الإطفاء لأن البيت الذي دخل إليه لم يحترق.

أجنحة المستشفى التي سمعت إنذارًا لم تسمعه الأجنحة الأخرى

بين أكتوبر 2019 ويوليو 2021، أجرت خمس مستشفيات تابعة للشؤون الصحية بوزارة الحرس الوطني في السعودية تجربة يصعب اختصار مغزاها في عبارة «نجحت التقنية». جرى توزيع خمسة وأربعين جناحًا طبيًا وجراحيًا ضمن تسع موجات انتقال بين الرعاية المعتادة ونظام إلكتروني للكشف المبكر عن الإنتان، أي الاستجابة المرضية الخطرة التي قد ترافق العدوى. لم تكن التجربة عن روبوت يشخّص المرض سرًا، ولا عن نموذج لغوي يكتب توصية علاجية. كانت أبسط، ولهذا كانت أكثر فائدة لسؤالنا.

يحسب السجل الطبي الإلكتروني إشارة إنذار اعتمادًا على ثلاثة مؤشرات من مقياس «كيو سوفا» (quick Sequential Organ Failure Assessment, qSOFA): انخفاض ضغط الدم الانقباضي إلى حد معين، أو ارتفاع معدل التنفس، أو اضطراب درجة الوعي. حين تظهر اثنتان من العلامات ضمن نافذة زمنية محددة، تُولد الإشارة. وفي الأجنحة التي انتقلت إلى التدخل صارت هذه الإشارات ظاهرة لفريق الرعاية ومتصلة بمسار استجابة، بينما استُخدم نمط غير معلن للمقارنة في الأجنحة الأخرى. هكذا أصبح بوسع الباحثين أن يسألوا سؤالًا أهم من دقة الحساب: ماذا تغيّر عندما أصبح لدى البشر إنذار يمكنهم التصرف بسببه؟1

شملت الدراسة 60,055 مريضًا، منهم 29,442 في مجموعة الفحص الإلكتروني و30,613 ضمن المقارنة. ازدادت احتمالات إجراء تحليل اللاكتات وطلب السوائل الوريدية بعد الإنذار في مجموعة التدخل. لكن قراءة جدول الوفيات دون سياقه تثير الحيرة: كانت النسبة الخام للوفاة داخل المستشفى خلال تسعين يومًا **3.2%** في التدخل مقابل **3.1%** في المقارنة، ولم يظهر فرق خام واضح لصالح الفحص. وبعد التحليل المحدد في الدراسة لضبط الفترات الزمنية وتجمع المرضى ضمن الأجنحة والمستشفيات وحالة كوفيد، قُدّر الخطر النسبي المعدّل بـ**0.85**، مع فاصل ثقة 95% من 0.77 إلى 0.93. وهذا تقدير لانخفاض نسبي في الخطر بعد الضبط، لا تصريح بأن كل مريض نجا بفضل البرنامج، ولا أن المقارنة الخام وحدها تكفي لقراءته.1

لِمَ هذا التفاوت بين الخام والمعدّل؟ لأن التجربة كانت ذات انتقال مرحلي؛ تغيّرت أجنحتها وأوقاتها، ثم وصلت جائحة كوفيد وغيّرت خصائص المرضى ومخاطرهم عبر الفترات. ذكر الباحثون أن تقييم الأثر احتاج إلى ضبط تلك العوامل. وهذه ملاحظة جوهرية في بحث عن النبوءات: حتى عندما نملك تجربة عشوائية، لا تُسلِّمنا الطبيعة جواباً سببيًا جاهزًا في صفٍّ واحد من الأرقام. نحتاج أن نعرف كيف عُرِّضت المجموعات للتدخل ومتى، وكيف تغيرت الظروف المحيطة بها.1

ليس هذا دليلًا على أن نموذج الذكاء الاصطناعي تنبأ بخطر ثم خسرت دقته بعد إنقاذ المرضى؛ فالنظام السعودي استخدم قاعدة qSOFA إلكترونية، ولم تختبر الورقة هذا الادعاء عن انحراف أداء نموذج تعلّم آلي بعد النشر. بل هو دليل تجريبي على الحلقة التي نتحدث عنها: **إشارة خطر تُغيّر أفعال الطاقم، وأفعال الطاقم يمكن أن تُغيّر الحدث الذي نحاول التنبؤ به.** ومن هنا تبدأ مشكلة مختلفة تمامًا عن جودة التصنيف.

ما الذي يعنيه أن يُخطئ إنذارٌ نجح؟

لنتصور مريضًا ظهرت عليه إشارات تدهور. يسجّل نظام الخطر احتمالية مرتفعة لحدث خطير خلال أربع وعشرين ساعة. يصل التنبيه إلى الفريق، فيُفحَص المريض سريعًا ويُقدّم التدخل الملائم، ثم لا يقع الحدث. بعد شهر تعود إدارة البيانات إلى السجل: «تنبّأ بالخطر ولم يحدث. أليس هذا إنذارًا كاذبًا؟».

قد يكون كذلك. وقد لا يكون.

إذا كانت النسبة التي قدمها النموذج تعني «احتمال الحدث **مع استمرار الرعاية المعتادة دون إضافة هذا الإنذار**»، فلا يصح امتحانها وكأنها تعني «احتمال الحدث **بعد أن استجبنا للإنذار**». فالمعلومة الجديدة غيّرت المسار الذي كان يُراد التنبؤ بنهايته. أما إذا صُمّم النموذج أصلًا لتوقع ما سيقع بعد الاستجابة المتوقعة من الطاقم، فإن مقارنته بالنتيجة الفعلية تحت تلك السياسة تصبح سؤالًا مشروعًا، لكن هدفه مختلف.

في علم الإحصاء قد نكتب خطر وقوع الحدث تحت السياسة المعتادة على الصورة: $R_0(x) = P(Y(0) = 1 \mid X = x)$. ونكتب الخطر تحت برنامج الإنذار والاستجابة: $R_1(x) = P(Y(1) = 1 \mid X = x)$. حيث ترمز X إلى المعلومات المعروفة قبل اتخاذ الفعل، و$Y(0)$ إلى النتيجة المحتملة في **الرعاية المعتادة دون البرنامج الجديد**، لا في عالم يخلو من الطب والعلاج؛ و$Y(1)$ إلى النتيجة المحتملة مع برنامج الإنذار. الفرق بينهما هو ما قد تُحدثه السياسة عند أشخاص من النوع نفسه، في ظل افتراضات تصميم محددة.

لا يُرى المآلان للمريض نفسه معًا. نلاحظ ما وقع في المسار الذي اتُّخذ، لا نسخة ثانية من العالم كان فيها المريض ذاته ولم يسمع الفريق الإنذار. تسمّى هذه مشكلة النتائج المحتملة المضادة للواقع (Potential Outcomes). وهي التي تجعل عبارة «لم يحدث شيء» أقل إفادة مما تبدو. لم يحدث شيء **بعد ماذا؟** وبعد أي تدخل؟ وما النتيجة التي يُرجح وقوعها لو استمر المسار الآخر؟

في 2020، قدّم خوان بيردومو (Juan Perdomo) وتيانا زرنيتش (Tijana Zrnic) وسيليستين مندلر-دونر (Celestine Mendler-Dünner) وموريتس هاردت (Moritz Hardt) إطار «التنبؤ الأدائي» (Performative Prediction). الفكرة أن النموذج حين يدعم قرارًا قد يؤثر في توزيع البيانات المستقبلية الذي سيُحاكم عليه، فلا يعود العالم الذي تنبأ به هو نفسه العالم الذي يعيش فيه بعد نشر توقعاته. تناول البحث خصائص رياضية للاستقرار بعد تكرار التدريب والتأثير المتبادل بين التنبؤ والبيانات، ولم يقدّم شهادة بأن كل إنذار تتراجع دقته قد نجح في الوقاية.2

الفرق بين النبوءة التي تغيّر سلوكنا وبين النبوءة التي لا تغيّر شيئًا ليس تفصيلًا فلسفيًا. توقع المطر لا يغيّر الغيوم في العادة، لكن توقع تدهور مريض قد يغيّر قرار الطبيب، وتوقع تعثر طالب قد يغيّر طريقة دعمه، وتوقع حادث صناعي قد يغيّر جدول الصيانة. هنا ليس الذكاء الاصطناعي مرآة للمستقبل، بل متغيراً جديدًا في صناعة المستقبل نفسه.

مستشفى استخرج إشارات الخطر من تصرفات الممرضات

نحتاج إلى حالة استخدمت تعلم الآلة فعلًا، لا مجرد قاعدة إنذار. في أبريل 2025، نشرت مجلة *Nature Medicine* تجربة عشوائية عن نظام **CONCERN** للإنذار المبكر بتدهور المرضى. لا يعتمد النظام أساسًا على التحاليل وحدها؛ إنه يتعلم من الأنماط الزمنية للتوثيق والمراقبة التمريضية: متى زادت الممرضة مرات القياس؟ متى دوّنت ملاحظة غير معتادة؟ ومتى أجرت مراقبة في توقيت لا يمليه الروتين؟ يصبح تغير سلوك الفريق نفسه إشارة تستبق بعض ما قد يتدهور في جسد المريض.3

أُجريت التجربة عبر **74 وحدة سريرية** موزعة عشوائيًا بين التدخل والرعاية المعتادة في نظامين صحيين أمريكيين، وشملت **60,893 إقامة بالمستشفى**. ارتبط إظهار إنذارات CONCERN ضمن سير العمل بانخفاض في الخطر اللحظي المعدّل للوفاة داخل المستشفى، مع نسبة مخاطر **0.64** وفاصل ثقة 95% من 0.53 إلى 0.78. لا تعني نسبة 0.64 أن «64% من المرضى تعافوا بسبب الذكاء الاصطناعي»؛ إنها نسبة معدل الخطر الآني المقارن في تحليل البقاء المستخدم بالدراسة. كما خُفضت مدة الإقامة في التحليل الوارد بالورقة.3

لكن في الجدول نفسه ارتفع الخطر اللحظي المعدّل لـ**التحويل غير المتوقع إلى العناية المركزة**، مع نسبة مخاطر **1.25** وفاصل ثقة من 1.09 إلى 1.43. كيف ينخفض الموت ويرتفع مؤشر يبدو للوهلة الأولى مؤشراً على التدهور؟ من الاحتمالات أن الرصد المبكر يدفع الفريق إلى نقل مزيد من المرضى إلى مستوى رعاية أعلى، لكن الدراسة لا تتيح لنا أن نحكم على كل تحويل بأنه إنقاذ ناجح. وقد يكون التحويل غير المتوقع في حالات أخرى نتيجة تدهور ينبغي منعه. نحتاج التوقيت والسبب والملاءمة السريرية وتغير المآلات، لا إشارة زائد أو ناقص أمام الرقم.3

هنا يصبح التقرير الإداري ذا وجهين. يمكن لقسم أن يقول إن النظام حسن أحد المآلات الخطرة، ويمكن لقسم آخر أن يلاحظ زيادة التحويلات غير المتوقعة، ويطلب تفسيرها. كلاهما يقرأ نتيجة صحيحة من الدراسة. لكن الفارق بين القراءتين يكشف حقيقة أكبر: **الإنذار الناجح قد يغيّر خريطة الأحداث، ولا يجوز أن نمنح أي حدث جديد معنى النجاح أو الفشل قبل أن نفهم مسار التدخل الذي أنتجه.**

وللدراسة قيود أعلنها الباحثون: أُجريت في نظامين صحيين أمريكيين وفي فترة تأثرت بكوفيد، مع اختلافات في مزيج وحدات العناية وممارسات التسجيل واحتمال تأثر بعض النتائج بعوامل لم تُقَس. لهذا تبقى دلالتها أقوى من قصة تسويقية، وأضيق من ضمان عام لكل مستشفى وكل نظام إنذار.3

في المدرسة أيضًا: انخفضت إشارات الخطر ولم يتحرك المؤشر الأخير

لئلا نحصر المفارقة في غرف العناية المركزة، فلننتقل إلى مدارس ثانوية أمريكية. في العام الدراسي 2014–2015 اختبرت دراسة أصدرتها وزارة التعليم الأمريكية لاحقًا نظام **EWIMS** للرصد والتدخل المبكر مع الطلاب المعرضين لخطر التعثر في التخرج. وُزعت **73 مدرسة ثانوية** عشوائيًا: 37 طبقت البرنامج و36 واصلت إجراءاتها المعتادة خلال سنة التقييم. شملت الدراسة **37,671 طالبًا** في السنتين الأولى والثانية من المرحلة الثانوية. لم يكن البرنامج نموذجًا لغويًا أو تجربة لنظام تعلّم آلي حديث؛ بل منظومة مؤشرات مبكرة ومتابعة وتدخل.4

بعد عام، بلغت نسبة التغيب المزمن **10%** في مدارس البرنامج مقابل **14%** في مدارس المقارنة. وبلغت نسبة الطلاب الذين رسبوا في مقرر واحد على الأقل **21%** مقابل **26%**. فروق ذات دلالة إحصائية في هذين المؤشرين. لكن نسبة الطلاب الذين لم يحصلوا على ساعات دراسية كافية ليكونوا على مسار التخرج خلال أربع سنوات كانت **14% في المجموعتين**؛ ولم تجد الدراسة أثرًا قابلًا للكشف على هذا المقياس للتقدم الدراسي بعد سنة واحدة. وأشارت كذلك إلى صعوبات تنفيذ الخطوات السبع للبرنامج وضعف التطبيق الكلي في كثير من المدارس.4

ما الدرس؟ قد يتغير السلوك الذي يصنع الخطر قبل أن يتغير المآل النهائي الذي نريده. الغياب المزمن إشارة، والرسوب في مقرر إشارة، والتقدم نحو التخرج نتيجة أقرب إلى الهدف البعيد. فإذا دربنا نظامًا جديدًا على سجلات لاحقة لتدخلات منعت بعض الغياب أو الرسوب، ثم عاملنا كل طالب تحسن بوصفه «تصنيف خطر خاطئًا»، خلطنا بين السؤال: «من كان يحتاج الدعم؟» والسؤال: «من تعثر رغم الدعم؟». لكن إن ارتفعت مؤشرات الحضور ولم يتحسن التقدم الدراسي، فلا يحق لنا بالمقابل إعلان أن الوقاية حققت هدفها النهائي.

المدرسة والمستشفى هنا لا يشتركان في التقنية أو تعريف النجاح. يشتركان في منطق واحد: عندما تُستخدم علامة الخطر لتغيير تصرف البشر، تصبح النتائج التالية جزءًا من **استجابة المؤسسة**، لا مجرد امتحان محايد لقدرة العلامة على التنبؤ.

لا يكفي أن تختار الإنسان المعرّض للخطر

هذا التمييز ضروري حتى لا يتحول الحديث عن «نجاح الإنذار الخفي» إلى عذر مريح لأي نموذج ضعيف. في مدينة كامدن بولاية نيوجيرسي الأمريكية، جرى اختبار برنامج مكثف لتنسيق الرعاية للمرضى المعروفين بكثرة دخول المستشفيات واحتياجاتهم المعقدة. كانت الفرضية التشغيلية جذابة: نعرف الأشخاص الذين يتكرر دخولهم؛ فلنوفر لهم دعمًا طبيًا واجتماعيًا منسقًا، ولتنخفض العودة إلى المستشفى.

لكن تجربة عشوائية شملت **800 مريض** لم تجد انخفاضًا ذا دلالة إحصائية في إعادة الدخول خلال 180 يومًا. كانت النسبة **62.3%** بين من تلقوا البرنامج مقابل **61.7%** في الرعاية المعتادة؛ وكان فرق التحليل المعدّل **0.82 نقطة مئوية** مع فاصل ثقة يمتد من انخفاض 5.97 نقاط إلى ارتفاع 7.61 نقطة. والأهم أن مقارنة المجموعة المتلقية للبرنامج بنفسها قبل التدخل وبعده كانت ستوحي بانخفاض كبير يقارب **38 نقطة مئوية**، لأن المرضى دخلوا التجربة بعد فترة من الاستخدام الاستثنائي المرتفع ثم انخفضت الزيارات في المجموعتين. لا يجوز نسب هذا الانخفاض تلقائيًا إلى البرنامج.[5]

لم يكن برنامج كامدن تجربة ذكاء اصطناعي. لكنه يقدم اعتراضًا تجريبيًا قاطع الأهمية على الخلط بين **العثور على الأكثر عرضة لحدث** و**العثور على من يمكن منع الحدث لديهم بالتدخل المتاح**. الشخص الأعلى خطورة ليس بالضرورة صاحب أكبر منفعة متوقعة من برنامج معين. قد تكون أسباب خطره لا تستجيب لهذا البرنامج، أو تتطلب موارد لا يقدمها، أو تكون مؤقتة بطبيعتها.

في الورق، يمكن لنموذجين أن يرتبا المرضى بحسب احتمال العودة إلى المستشفى، فيتفقا على مَن هو الأكثر عرضة. لكن لو سألنا أي المرضى سيستفيد أكثر من تدخل محدد، فقد نحتاج ترتيبًا مختلفًا تمامًا. أحد السؤالين عن **مستوى الخطر**، والآخر عن **قابلية الخطر للتغيير**. والخلط بينهما يجعل دقة تنبؤية عالية تبدو كأنها سياسة وقائية ممتازة دون دليل.

تظهر العقدة نفسها في دراسة منشورة في فبراير 2026 عن عرض سلبي لتنبؤات الذكاء الاصطناعي بتدهور المرضى في جناح قلبي، شملت **10,422 إقامة** وجرّبت إظهار مسارات الخطر دون فرض استجابة سريرية محددة. لم تجد الدراسة تحسنًا في المخرج الأساسي. ونقل الأطباء نحو **11%** من المرضى بين أسرّة العرض وعدم العرض، بما أضعف بعض خصائص العشوائية العملية. ليست النتيجة حكمًا بفشل كل نماذج الإنذار، لكنها تذكير بأن **معرفة الخطر لا تُنتج تلقائيًا فعلًا قادرًا على تقليله**.[6]

بهذا يصبح لدينا شقان لا يصح فصل أحدهما عن الآخر: ربما كان الإنذار صحيحًا وأنقذ التدخل حالة، وربما كان الإنذار غير دقيق، أو كان التدخل بلا أثر. ما يفصل الاحتمالات تصميم سببي، لا حسن الظن بالخوارزمية.

النتيجة التي نحتاجها ليست النتيجة التي نملكها

في ورقة بحثية نشرت عام 2020، شددت نان فان خيلوفن (Nan van Geloven) وزملاؤها على ضرورة تعريف الشيء الذي يتنبأ به النموذج عندما يبدأ العلاج بعد لحظة قياس الخطر. أطلقوا على هذا التحديد اسم **Predictimand**، على غرار تحديد مقدار الأثر المستهدف في التجارب السريرية. هل نتنبأ بما سيحدث إذا لم يبدأ علاج إضافي؟ أم بما سيحدث مع

استمرار الممارسات العلاجية كما كانت في بيانات التدريب؟ أم بوقوع الحدث قبل بدء العلاج؟ هذه أسئلة ذات إجابات مختلفة، وطرق تقدير مختلفة، وافتراضات مختلفة.7

وفي أكتوبر 2025، قدم كريستوفر بوير (Christopher B. Boyer) وزملاؤه عملًا أكثر مباشرة في **تقدير النماذج التنبؤية المضادة للواقع وتقييمها.** أوضحوا لماذا لا يكفي قياس جودة نموذج يتنبأ بنتيجة تحت سياسة علاج افتراضية باستخدام البيانات المرصودة تحت سياسة مختلفة، واقترحوا طرقًا لتقدير خصائص مثل الخسارة والمعايرة والمساحة تحت منحنى ROC في العالم المضاد للواقع، مع بيان شروط التعرف الإحصائي وافتراضاته. مثّلوا ذلك بنموذج خطر أمراض القلب عند من لم يبدأوا علاج الستاتين. هذه نتائج منهجية مشروطة، وليست وصفة تجعل المآلات غير المرصودة قابلة للمشاهدة.8

وهنا يجب أن نفرق بين ثلاثة مقادير:

1. **خطر الأساس:** احتمال وقوع الضرر في ظل السياسة المرجعية التي سنقارن بها، مثل الرعاية المعتادة دون نظام الإنذار الجديد.
2. **خطر ما بعد التدخل:** احتمال وقوع الضرر عندما يُستخدم التنبيه وفق سياسة معلومة، وبالاستجابة الفعلية التي يتبعها الناس.
3. **منفعة السياسة:** مقدار الضرر الذي أُنقص بفضل الانتقال من السياسة الأولى إلى الثانية، بعد احتساب الأعراض الجانبية والتكلفة والموارد والآثار الأخرى.

قد يكون خطر الأساس مرتفعًا، وخطر ما بعد التدخل منخفضًا، ومنفعة السياسة كبيرة. وقد يكون خطر الأساس مرتفعًا أيضًا لكن التدخل لا ينفع، فلا تتولد منفعة تُذكر. وقد يكون التدخل مؤذيًا لبعض الفئات. لا نستطيع أن نحدد الاختيار الصحيح من «درجة خطر» واحدة.

للتوضيح العددي فقط، تخيل شخصًا أ يبلغ خطر الحدث لديه 30% تحت السياسة المعتادة، وينخفض إلى 10% تحت تدخل معين. وتخيل شخصًا ب يبلغ خطره 20% وينخفض إلى 19%. الفرق التقديري في المنفعة هو 20 نقطة مئوية للأول ونقطة واحدة للثاني. هذه **أرقام افتراضية للتفسير وليست نتائج دراسة**؛ ولا يمكن معرفة المنفعة الفردية المتحققة من مشاهدة نتيجة كل شخص مرة واحدة. إنما نستطيع محاولة تقدير متوسطات آثار للفئات تحت افتراضات وتصاميم بحثية ملائمة.

تزداد المشكلة حين يعيد النظام تدريب نفسه دوريًا على البيانات التي نتجت عن استخدامه. فعندما يرى النموذج أن أشخاصًا منحهم درجات عالية لم يصابوا بالحدث بعد أن عولجوا أو نُقلوا أو دُعموا، قد يتعلم علاقة جديدة بين خصائصهم والنتيجة المرصودة. وقد يصبح أفضل في توقع ما سيحدث **تحت السياسة الحالية**، وأقل ملاءمة لتقدير ما كان سيحدث **لو لم تُطبّق تلك السياسة.** إعادة التدريب ليست خطأ دائمًا؛ لكنها تستلزم ثبات تعريف السؤال. لا يصح أن نغير الشيء المتنبأ به بصمت ثم نسمي ذلك مجرد تحديث للإصدار.

لماذا تُضلّلنا درجة الدقة؟

إذا أعلنت شركة أن نموذجها بلغت مساحته تحت منحنى ROC مقدارًا مرتفعًا، فهذا قد يثبت قدرة على التمييز في مجموعة اختبار محددة، لكنه لا يجيب وحده عن أثر الاستخدام على المرضى أو الطلاب أو المنشآت. وإذا انخفض مقياس التنبؤ بعد نشر تدخل ناجح، فلا يصح تفسير الانخفاض بلا معرفة الهدف الأصلي وطبيعة السياسة. ربما تغير توزيع الأحداث بفعل التدخل، وربما تغيّرت خصائص السكان، وربما تراجعت جودة النموذج فعلًا.

هناك ثلاث محاكمات مختلفة يجب ألا تبتلع إحداها الأخريين:

الأولى، محاكمة التوقع: هل كانت تقديرات الخطر سليمة في العالم المرجعي المحدد؟ يشمل ذلك المعايرة (Calibration)، والتمييز (Discrimination)، وصحة البيانات، وملاءمة الفترة والسكان.

الثانية، محاكمة الاستجابة: هل أحدثت الرسالة تغييرًا في سلوك صاحب القرار؟ كم استغرق من الإنذار إلى الفحص أو الإجراء؟ هل كان الفعل مناسبًا؟ وهل استجابت له الفئة المعنية؟

الثالثة، محاكمة النتيجة: ما أثر السياسة مقارنة برعاية معتادة أو سياسة بديلة؟ هل انخفض الضرر؟ وهل ظهرت أضرار أخرى، أو تكاليف، أو تحويلات غير ضرورية، أو استنزاف للعاملين والموارد؟

لا يمكن اختزال هذه المحاكمات في نسبة واحدة. في التجربة السعودية كانت زيادة الفحوص وطلبات السوائل جزءًا من استجابة الطاقم، وكانت الوفيات مخرجًا منفصلًا. وفي CONCERN انخفض أحد الأخطار بينما ارتفعت تحويلات العناية المركزة. وفي مدارس EWIMS تحسنت بعض المؤشرات دون أن يظهر أثر على تقدم التخرج بعد عام. وفي كامدن لم يثبت أن البرنامج حقق منفعة على الهدف الأساسي رغم استهداف الفئة ذات الاستخدام المرتفع. تختلف الآليات، لكن الحكم العلمي في جميع الحالات يتطلب الفصل بين التنبؤ والفعل والأثر.

والمشكلة هنا مختلفة عن الحالات التي تنعدم فيها الملصقات لأن بعض الأشخاص لم تُتَح لهم فرصة إظهار النتيجة. ففي التنبؤ الوقائي قد تكون النتيجة **مرصودة تمامًا**، لكنها نتيجة حدثت **بعد أن تدخلت السياسة.** غياب الحدث في السجل لا يكشف وحده إن كان سينشأ عند غياب التدخل. هذا ليس نقصًا في حقل داخل الجدول؛ إنه غياب مسار بديل من العالم.

كيف نختبر نموذجًا يمنع بعض ما يتوقعه؟

هناك طريق تجريبي مباشر حين يكون ممكنًا أخلاقيًا ومؤسسيًا: نحافظ على الرعاية أو الخدمة المعتادة للجميع، ونوزع **إضافة البرنامج الجديد** عشوائيًا بين وحدات متقاربة أو على مراحل مدروسة، ونسجّل التوقعات قبل كشفها، وتوقيت كشفها، وأفعال العاملين، والنتائج اللاحقة. لا تعني المقارنة منع علاج واجب أو تعريض فرد لخطر معروف. يتعلق الاختبار بالميزة الإضافية التي لم تثبت فائدتها بعد. ويجب أن يقره أصحاب الاختصاص والحوكمة الأخلاقية في الحالات الحساسة.

يمكن لبعض الأنظمة إبقاء درجة الخطر محسوبة في وضع غير ظاهر أثناء المقارنة، على أن تُحمى المعلومات وتُحترم حدود الرعاية المعتادة. ويُقاس أثر **إتاحة الإنذار** على مستوى المجموعات أو الأجنحة، لا بأن ندّعي معرفة من كان سيموت بين من عاشوا. وتجارب الانتقال المرحلي والعشوائية العنقودية مفيدة حين تتسرب المعرفة بين العاملين في الوحدة نفسها، لكن لها تحديات الفترات والأحداث العامة، كما برهنت تجربة كوفيد المصاحبة للمستشفيات السعودية.

عندما لا يمكن إجراء تجربة عشوائية، يمكن محاولة محاكاة تجربة هدف (Target Trial Emulation) من بيانات رصدية، أو استخدام طرق ضبط التداخل العلاجي وتقدير الأثر المضاد للواقع. غير أن هذه الطرق تعتمد على افتراضات لا ينبغي إخفاؤها: معرفة عوامل الالتباس المؤثرة في القرار والنتيجة، وإمكان مقارنة أشخاص لديهم فرص معقولة للتعرض لخيارات مختلفة، وتعريف التدخل بطريقة متسقة، وعدم إفساد المقارنة بتأثيرات متبادلة غير معالجة بين الوحدات. قد لا تصح الافتراضات في كل بيئة.

وإذا كان تدخل شخص واحد يستهلك سرير عناية يحتاجه آخر، فالمسألة لا تقف عند أثر فرد معزول. نحن أمام تخصيص موارد وازدحام وتأثيرات متبادلة (Interference). قد تنفع سياسة عند عشرات المرضى وتختلف نتيجتها عند تطبيقها على كل أجنحة البلد. ولهذا لا يكفي اختبار دقة النموذج ثم افتراض أن مضاعفة التنبيهات تضاعف الإنقاذ. تختلف قدرة الاستجابة، وتظهر كلفة الإنذارات المفرطة، وقد تتغير سلوكيات الفرق نفسها مع مرور الوقت.

دفتر التنبؤ والتدخل: اقتراح قابل للدحض

أقترح أن تحتفظ المؤسسة بـ«**دفتر التنبؤ والتدخل**» (Forecast–Intervention Ledger)، وهي مساهمة تحريرية مقترحة وليست معيارًا معتمدًا أو نظامًا مثبت الأداء. لا يُنشأ الدفتر لتكثير الحقول في لوحة التحكم، بل لفصل ثلاثة أشياء كانت تختلط: ما عرفناه قبل التدخل، وما فعلناه بسببه، وما الذي حدث بعده.

لكل حالة مؤهلة، يدوّن زمن جمع المعلومات السابقة، ونسخة النموذج، وتعريف الخطر الذي تتنبأ به الدرجة، وأفق التنبؤ، ووقت إصدار الإنذار، وهل ظهر لمتخذ القرار، والاستجابة التي أعقبته وتوقيتها، ونتائج الحالة، وأضرارها الممكنة. ثم يوجد مستوى تحليلي مستقل يقدّر الأثر السببي على مجموعات قابلة للمقارنة، ويحفظ هامش عدم اليقين وافتراضاته. لا يُملأ حقل «ما كان سيحدث لهذا المريض لو لم يُنذر» برقم فردي مختلق؛ لأن ذلك المستقبل لم يُشاهد.

يمكن تحويل هذا الدفتر إلى ثلاث لوحات تراجعها المؤسسة منفصلة:

- **لوحة التنبؤ:** معايرة النماذج تحت السياسة المرجعية، وحساسية الإنذار، وملاءمة درجة المخاطر للمجموعة المستهدفة.
- **لوحة التدخل:** نسبة الإنذارات التي نتجت عنها استجابة مناسبة، والفواصل الزمنية، والأعباء والإحالات والأضرار.
- **لوحة الأثر:** فرق النتائج مقابل سياسة مقارنة محددة، مع فواصل الثقة والتكلفة وأثر توزيع الموارد، وليس مجرد ما حدث للمُنذَرين.

وتكون التجربة التي تدحض قيمة النظام واضحة: إذا لم يظهر فرق معتبر في النتيجة المهمة تحت مقارنة جيدة، أو كانت أضرار الإنذارات تفوق المنافع، أو كان تحديد الفئة الأكثر خطورة لا يغيّر جدوى التدخل، فلا تنقذه بلاغة «لقد منعنا أحداثًا لا تُرى». بل يجب أن يتحمل النموذج احتمال أنه لم يحسّن العالم، كما تحملنا احتمال نجاحه قبل أن تتكلم النتائج.

أهم ما في الدفتر أن الإجابة لا تكون أبدًا: «انخفض عدد الحالات بعد تركيب النظام، إذن أنقذها». فقد تنخفض الحالات بسبب تغير السكان أو الموسم أو الرعاية أو انحسار ذروة سابقة. درس كامدن في هذا الباب بالغ القسوة: مقارنة ما قبل البرنامج وما بعده أوحت بتحسن كبير، لكن المجموعة العشوائية المقارنة شهدت مسارًا مشابهًا.

الاختبار الذي يهم مجلس إدارة المستشفى، أو وزارة التعليم، أو جهازًا مسؤولًا عن السلامة العامة، ليس عدد التنبيهات التي أصدرها النظام ولا عدد الأيام التي لم تقع فيها الكارثة وحده. إنه: **ما مقدار الضرر الذي تغير بسبب هذه السياسة تحديدًا، ولمن، وبأي تكلفة، وبأي ثقة في الاستنتاج؟**

حين يصبح المستقبل الثاني جزءًا من المساءلة العلمية

ليس مطلوبًا من الذكاء الاصطناعي أن يعرف الغيب، ولا من الباحث أن يعرض مستقبلًا لم يحدث على شاشة. المطلوب أن يعترف بأن القرارات الوقائية تخلق مفارقة لا تحلها اختبارات التصنيف التقليدية وحدها: قد تكون بعض النتائج السليمة ثمرة تدخل ناجح، وبعضها الآخر ثمرة إنذار زائد أو حظ أو تغير طبيعي. والخوارزمية لا تستطيع أن تمنحنا الفرق باختيار الرواية الألطف عن نفسها.

لذلك لا ينبغي أن نكافئ النموذج لأنه خاف من كل شيء، ولا أن نعاقبه لمجرد أن ما خاف منه لم يحدث. لا ينبغي أن نخلط خطر المريض بقابلية علاجه، ولا تعثر الطالب باحتمال تحسنه حين يجد معلمًا يساعده، ولا ننسب كل انخفاض بعد برنامج إلى البرنامج.

أعمق تحول يفرضه التنبؤ الأدائي على المؤسسة هو الانتقال من سؤال يبدو علميًا: «هل توقعت ما سيحدث؟» إلى سؤال أشد صعوبة وأشد إنصافًا: «ماذا كان سيحدث لولا أن تصرفنا بناءً على توقعك، وكيف نستطيع اختبار الفرق دون ادعاء أننا رأينا المستقبلين؟».

عند هذه النقطة يتغير معنى النجاح. لن تكون أفضل آلة إنذار هي التي يتحقق أكبر عدد من تحذيراتها، لأن تحقق الخطر ليس غاية الوقاية. ولن تكون الأفضل بالضرورة هي التي تختفي بعدها كل الأحداث، لأن اختفاء الحدث قد لا يكون من صنعها. ستكون الأفضل هي التي نستطيع أن نثبت، بتجربة أو باستدلال سببي منضبط، أنها ساعدت البشر على جعل الاحتمال السيئ أقل وقوعًا، من دون أن تخفي الأثمان والأضرار البديلة.

لقد تعلمنا كيف نختبر ما إذا كان الحاسوب يعرف شيئًا عن المستقبل. بقي أن نتعلم كيف نختبر ما إذا كانت **معرفته قد جعلت المستقبل أفضل.**

المراجع

1. Arabi et al., JAMA, Electronic Sepsis Screening (2025 print; 2024 online).
2. Perdomo et al., Performative Prediction, ICML 2020.
3. Rossetti et al., CONCERN RCT, Nature Medicine 2025.
4. Institute of Education Sciences, EWIMS Impact Study, April 2017.
5. Finkelstein et al., Health Care Hotspotting RCT, NEJM 2020.
6. Keim-Malpass et al., AI-based Analytics RCT, Scientific Reports 2026.
7. van Geloven et al., Predictimands, European Journal of Epidemiology 2020.
8. Boyer et al., Counterfactual Prediction Evaluation, Statistics in Medicine 2025.