

5. La ciència de dades i l'exposoma

El principal avantatge de l'exposoma respecte als enfocaments tradicionals d'estudi «una exposició - una malaltia» és que proporciona un marc conceptual sense precedents per a l'estudi de múltiples perills ambientals (urbans, químics, estils de vida, socials) i els efectes que van associats amb ells. De fet, els models clàssics d'un sol contaminant no asseguren que l'associació analitzada sigui deguda a l'efecte contaminant o a una altra exposició correlacionada que no s'ha tingut en compte en l'anàlisi. Tampoc són capaços de capturar les interaccions i els efectes acumulatius de la barreja d'exposició. A més, atès l'augment de disponibilitat de dades complexes de salut ambiental a causa de l'aparició de noves tecnologies (com les històries clíniques electròniques, plataformes òmiques d'alt rendiment, sensors portàtils, etc.), calen enfocaments estadístics més avançats que se centrin en mescles complexes d'exposicions.

No obstant això, l'anàlisi de dades tan complexes comporta nombrosos reptes, per exemple, les correlacions generalment altes entre les exposicions d'una mateixa família (contaminants atmosfèrics, estil de vida), la capacitat de capturar efectes acumulats de dosis baixes, avaluar les interaccions o identificar components importants de la mescla. Recentment, s'han aplicat mètodes que tinguin en compte múltiples exposicions i les interaccions entre elles, per exemple, mitjançant l'ús de mètodes d'anàlisi de mescles, integrant la selecció, la contracció i l'agrupació de variables correlacionades (per exemple, LASSO, *elastic-net*, *adaptive elastic-net*), tècniques de reducció de dimensions (per exemple, component principal, anàlisi de mínims quadrats parcials) o mitjana de model bayesià (BMA), regressió de màquina del nucli bayesià (BKMR), etc. Una sèrie de limitacions d'aquests enfocaments s'han identificat prèviament, com ara la manca d'estabilitat de selecció de models (mètodes de contracció), la manca d'interpretabilitat de les variables latents (reducció de dimensions) i la ineficiència compu-

tacional (models bayesians). A més, poques vegades s'apliquen en el context de dades de l'exposoma grans (>100 variables) i heterogènies (òmiques, variables categòriques/contínues).

Reducció de la dimensionalitat

El primer punt d'interès és l'ampli ús de mètodes de reducció de la dimensionalitat, especialment la selecció de característiques, per a tractar dades exposòmiques multivariants. Els mètodes d'extracció de característiques són menys utilitzats perquè solen complicar la interpretació dels resultats si ens interessa l'efecte d'una determinada exposició sobre la salut. No obstant això, alguns mètodes interessants intenten analitzar grups d'exposicions correlacionades com una manera de reduir la dimensionalitat de l'entrada mantenint els resultats interpretables.

Efectes combinats de les exposicions

Per a mesurar l'efecte combinat de les exposicions es poden utilitzar mètodes d'índex. Aquests tenen l'avantatge d'una interpretació fàcil, ja que proporcionen una estimació d'un únic paràmetre (és a dir, un únic valor numèric) per a la barreja d'exposicions, juntament amb les ponderacions que il·lustren fàcilment la contribució de cada exposició. Tots els mètodes d'índex tenen l'inconvenient que no tenen en compte les interaccions entre exposicions que contribueixen al mateix índex. Això es pot resoldre utilitzant mètodes de superfície de resposta, a costa d'una interpretació potencialment més difícil. La tensió entre interpretabilitat i complexitat a l'hora d'escollir entre models d'índex i models de superfície de resposta es pot alleujar mitjançant mètodes desenvolupats recentment (models d'índexs múltiples) que combinen alguns dels avantatges d'ambdues famílies de mètodes. Aquests mètodes defineixen índexs fàcilment interpretables, i, a més, poden incorporar relacions no lineals i no additives entre els índexs d'exposició i el resultat de salut (McGee *et al.*, 2023).

Les tècniques bayesianes¹ també poden ser útils perquè penalitzen de manera natural els models complexos i ofereixen flexibilitat per a incorporar un procés de selecció variable. També és útil obtenir les distribucions de qualsevol quantitat que es pugui derivar de la sortida del model.

1. Les tècniques bayesianes són mètodes estadístics que impliquen l'actualització de creences o probabilitats sobre hipòtesis basades en coneixements previs i dades observades, i permeten la incorporació de la incertesa i l'estimació de paràmetres mitjançant distribucions de probabilitat.

Aprenentatge automàtic (machine learning) i predicció

Els mètodes d'aprenentatge automàtic poden augmentar potencialment la predictibilitat del resultat capturant informació més complexa de les dades (per exemple, interaccions complexes, relacions no lineals, etc.). Els models que combinen múltiples tècniques estadístiques en un conjunt poden fins i tot proporcionar millors resultats, ja que els diferents mètodes poden ser capaços de capturar diversos patrons de dades. Els mètodes d'aprenentatge automàtic inclouen mètodes de conjunt (com ara boscos aleatoris, *xgboost*), xarxes neuronals i màquines de vectors de suport.

Models causals

Els models causals han guanyat popularitat en epidemiologia ambiental (Bind, 2019), fins i tot per a mescles. De fet, les qüestions causals són les que en última instància impulsen les intervencions i el canvi de polítiques. L'anàlisi de la mediació causal amb dades d'exposoma ens pot ajudar a prioritzar els factors ambientals que tenen un impacte més gran en la salut. Alguns exemples de models causals aplicats a les associacions exposoma-salut inclouen l'anàlisi de mediació mitjançant dades òmiques, mètodes de computació-g (*g-computation*) i l'ús de boscos aleatoris causals (*causal random forest*).