Universitat Autònoma de Barcelona

Predicció anatòmico-clínica del risc d'evolució desfavorable en el pacient amb trastorn mental

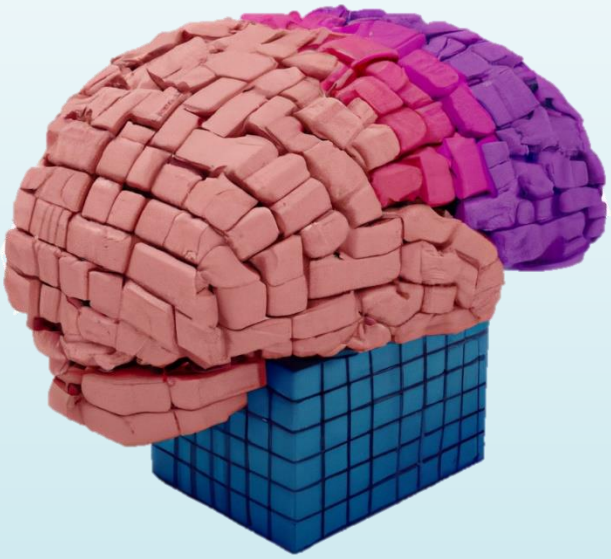
ALEIX SOLANES FONT

Predicció anatòmico-clínica del risc d'evolució desfavorable en el pacient amb trastorn mental

cap a la medicina personalitzada

PER

ALEIX SOLANES FONT



Tesi doctoral

Programa de Doctorat en Psiquiatria

Departament de Psiquiatria i Medicina Legal

Universitat Autònoma de Barcelona

Universitat Autònoma de Barcelona

Barcelona, Gener de 2023

Dirigida per:

Joaquim Raduà Castaño

Eduard Vieta Pascual

Tutoritzada i dirigida per:

Albert Fernández Teruel

CERTIFICAT DE DIRECCIÓ

Universitat Autònoma de Barcelona

Certificat de direcció de tesi doctoral

Dr. Joaquim Raduà Castaño, cap del grup d'Imatge i Trastorns relacionats amb l'estat d'ànim i l'ansietat de l'Institut d'Investigacions Biomèdiques August Pi i Sunyer.

Certifico:

Que el treball de tesi amb títol: "*Predicció anatòmico-clínica del risc d'evolució desfavorable en el pacient amb trastorn mental*" del programa de doctorat del Departament de Psiquiatria i Medicina Legal realitzat pel doctorand **Aleix Solanes Font** ha estat dirigit per mi.

I, perquè consti, signo aquest certificat.

Barcelona, 17 de gener de 2023

Universitat Autònoma de Barcelona

Certificat de direcció de tesi doctoral

Dr. Eduard Vieta Pascual, cap del servei de psiquiatria i psicologia de l'Hospital Clínic de Barcelona i líder del grup de recerca de Trastorns bipolars i depressius de l'Institut d'Investigacions Biomèdiques August Pi i Sunyer.

Certifico:

Que el treball de tesi amb títol: "*Predicció anatòmico-clínica del risc d'evolució desfavorable en el pacient amb trastorn mental*" del programa de doctorat del Departament de Psiquiatria i Medicina Legal realitzat pel doctorand **Aleix Solanes Font** ha estat dirigit per mi.

I, perquè consti, signo aquest certificat.

Barcelona, 17 de gener de 2023

Universitat Autònoma de Barcelona

Certificat de direcció de tesi doctoral

Dr. Albert Fernández Teruel, professor del Departament de Psiquiatria i Medicina Legal de la Unversitat Autònoma de Barcelona.

Certifico:

Que el treball de tesi amb títol: "*Predicció anatòmico-clínica del risc d'evolució desfavorable en el pacient amb trastorn mental*" del programa de doctorat en psiquiatria del Departament de Psiquiatria i Medicina Legal realitzat pel doctorand **Aleix Solanes Font** ha estat tutoritzat i dirigit per mi.

I, perquè consti, signo aquest certificat.

Barcelona, 17 de gener de 2023

AGRAÏMENTS

Probablement, si m'haguessin dit una d'aquelles nits de carregar instruments i voltar pels pobles de Catalunya a fer concerts, que un dia em tocaria escriure uns agraïments per una tesi doctoral, no crec que ho hagués cregut. No ha entrat mai dins els meus plans vitals, però mai se sap quins trencalls pots anar trobant darrere de cada obstacle, i el més xocant, és que realment tot agafa sentit quan ho mires amb perspectiva.

Vull començar els agraïments per les dues persones que em van donar la benvinguda a l'estrany món de la recerca, el Quim i l'Edith. Encara recordo la primera reunió on em parlaven del projecte sobre una cosa anomenada "neuroimatge", la qual desconeixia i que sonava quasi a ciència-ficció. Gràcies a ells (o per culpa seva) estic aquí en aquesta etapa tan curiosa, sense oblidar la Laura Igual que després de moltes reunions i estones de parlar sobre tècniques diverses d'Intel·ligència Artificial i Matemàtiques va ser qui em va proposar per col·laborar en aquest nou projecte.

Al Quim, a part, li he d'agrair tota la confiança, la paciència, l'ajuda, i totes les hores i hores que hem compartit debatent i estructurant tot el que ha acabat sent aquesta tesi. Les classes privades d'estadística, les lliçons sobre què és o hauria de ser la recerca, i també la seva vessant sempre positiva de tot plegat. Sempre hi ha algun article o alguna dada que pugui fer que la vida sigui bonica.

Voldria agrair també a un altre membre de la meva ja família de recerca, l'Anton, qui em va descobrir un món nou en massa sentits (MASSA). Sempre a punt per rebatre qualsevol idea, però sempre amb un somriure i alguna idea política de fons que et fa veure que un món ideal potser no és possible, però que sempre hi ha motius per seguir creient que tot pot millorar. No parlaré dels five-fingers, el goat...(no ho acabo per si de cas) o del despacito a Vancouver.

Al Miquel Àngel, amb qui sempre és un plaer compartir converses sobre qualsevol tema, o sobre bàsquet (encara que no sigui de l'equip correcte). Sempre ha estat disponible per tot, fins i tot en el seu moment va ser tutor d'aquesta tesi! Entre ell i el Quim crec que m'han ensenyat més sobre el món de la recerca que tot el que hagi pogut aprendre de qualsevol altra banda.

Al Daniel Vega i l'Albert Fernández, per la seva ajuda i disponibilitat com a tutors i al Dr. Antoni Bulbena pels seus consells i l'ajuda durant tot aquest període.

Seguint pel meu camí dins aquest món també vull agrair al Rai per totes les converses, discussions estadístiques junt amb el Quim, o les frases enginyoses que després de la rialla sempre et fan pensar. A l'Erick, amb qui les converses sobre com ens va la vida, o les vegades que he parlat sobre l'estat de la tesi no es poden ni comptar, i amb qui sempre hi guardaré una amistat especial per cada hora que vam compartir a FIDMAG. I parlant de FIDMAG no em vull oblidar de tots els amics que encara hi conservo, el Salva, la Mar amb la seva troupe, i tota la colla que hi vaig conèixer.

I seguint pel recorregut, vull agrair a l'Eduard per l'acollida que ens va brindar quan ens vam mudar al Clínic. Una persona que amb un caràcter tan proper i sempre amb una broma apunt és d'agrair, especialment de persones amb la seva carrera. Seguint pel clínic, també agrair a tot el grup de trastorns bipolars i depressius, la calidesa i simpatia que heu mostrat sempre. A l'Íria, per sempre tenir unes paraules amables i per l'interès i consells especialment en la nostra arribada a l'IDIBAPS i que encara segueix, al Jose per la simpatia,.. i un llarg etc. Perquè si una hi ha en aquesta unitat, és una família àmplia.

Per acabar amb el món de la recerca, vull agrair a tota una sala que ha fet que el dia a dia a l'IDIBAPS sigui un lloc ple de llum tot i estar en un soterrani sense ni una finestra. A l'Eloy (amb les inacabables discussions sobre qualsevol cosa, o buscant les coses que fan que el món pugui ser un món millor), el Carlos (amb les primeres converses sempre agradables per començar el dia), la Lydia (gràcies per la teràpia i els ànims quan ha fet falta), el Jose i la seva simpatia, la Maria, l'Enric que tot just comença, i a tots i cadascun dels que dia a dia compartiu estones per fer el dia més agradable, ja sabeu qui sou.

Per últim, a les persones més importants de totes, aquelles que m'han vist evolucionar (crec que ho he fet, tot i que no estic segur si cap a bé), als meus pares Josep i Fina, perquè m'heu recolzat en absolutament tot el que he volgut fer, i quan més ho he necessitat sempre hi heu estat. Aquesta tesi és en bona part culpa vostra! Al meu germà Xavi, que sempre ha estat a punt per burxar o fer broma, tot i que sé que en el fons sempre quan li he demanat consell, ajuda, o qualsevol cosa, sempre m'ha mostrat la seva estima.

I finalment, a la persona sense la qual estic segur que no estaria on estic, gràcies, Carla. Sempre ha estat una persona que m'ha fet costat, m'ha ajudat en les etapes més difícils de la meva vida, i va fer que de cop i volta, dels problemes inicials acabés on estic ara, a punt de presentar una tesi doctoral. La persona a qui admiro més, i per molts motius. Sense el teu suport res seria possible.

Sé que probablement m'he descuidat persones que mereixerien estar aquí, no m'ho tingueu en compte, si m'heu fet riure alguna vegada, el vostre nom en el fons també hi és present.

Bonus: Gràcies a la universitat per l'ajuda, la claredat i les facilitats que ha posat en tot moment.

SUMARI

LLISTA DE FIGURES

PREFACI

El meu interès per la recerca en neuroimatge aplicada als trastorns mentals va aparèixer en acabar el màster en Intel·ligència Artificial (IA). Tenia clara la gran potencialitat de les tècniques en IA i també que allò al qual volia dedicar-me fos algun camp de la IA que pogués repercutir en millorar la vida de les persones. No obstant això, ningú em va advertir com de difícil pot ser treballar amb dades de neuroimatge cerebral. Això fa que sigui un repte diari intentar entendre què podria explicar-nos el cervell respecte la nostra salut.

La recerca en salut mental no s'entén ja sense un enfocament multidisciplinari. La combinació de diferents disciplines com ara psicòlegs, psiquiatres, físics, enginyers de camps diversos o informàtics poden donar una comprensió més completa i efectiva als trastorns mentals. Les diferents experiències personals amb familiars o coneguts que han viscut algun problema de salut mental en algun moment de la seva vida han servit per fer-me veure que no hi pot haver res més valuós que intentar ajudar perquè en un futur els trastorns mentals puguin ser més fàcilment tractables o predictibles.

Aquesta tesi és el meu primer pas per intentar contribuir a fer que cada dia un nombre menor de persones hagi d'enfrontar-se a les dificultats derivades d'un trastorn mental. Espero que aquelles persones que la llegiu, hi trobeu alguna cosa que us faci gaudir-la ni que sigui la meitat del que m'ha costat fer-la.

1. INTRODUCCIÓ

Els tres objectius principals en la teràpia en pacients amb trastorn mental són aconseguir una remissió simptomàtica, la recuperació social i reduir el risc de futures recaigudes. Poder estimar el risc que succeeixin aquests objectius serà, per tant, de màxima importància. Crear eines que permetin realitzar aquesta estimació podrà ajudar de moltes formes diferents als pacients: ajustar els fàrmacs implicaria als pacients amb major risc de pobra remissió simptomàtica poder beneficiar-se d'intervencions psicològiques [1], als pacients amb major risc de disfunció social rebre tractaments de rehabilitació funcional [2], als pacients amb risc de discontinuació de la medicació ser inclosos en iniciatives destinades a reduir la no adherència al tractament [3], o als pacients amb risc de recaiguda depressiva poder beneficiar-se de teràpia cognitiva-conductual, teràpia centrada en la família o psicoeducació [4]. I si bé, és important millorar els símptomes de l'episodi actual per tal d'aconseguir la remissió simptomàtica, és igualment important preveure el risc de recaiguda, ja que al voltant del 30% dels pacients amb un primer brot psicòtic recauran abans del primer any, i fins a un 80% ho faran en els propers 5 anys[5].

Tradicionalment, la majoria dels estudis de neuroimatge en psiquiatria s'han basat en enfocaments estadístics univariats. Per exemple, tècniques com la morfometria basada en vòxel (VBM) permeten avaluar les diferències entre pacients i controls en el volum regional o la composició tissular a partir d'estimacions de probabilitat de teixits. Aquests enfocaments són útils per detectar diferències de grup com poden ser les anomalies associades a certs trastorns. Així i tot, no permeten fer prediccions individualitzades (com ara de l'evolució de la malaltia). En els últims anys, l'interès pels mètodes d'intel·ligència artificial en neuroimatge ha augmentat a causa de la seva capacitat per manejar dades amb una alta dimensionalitat i de realitzar prediccions a nivell individual.

L'ús d'algorismes d'aprenentatge automàtic com les màquines de vectors de suport (Support Vector Machines (SVM)) [6], la regressió regularitzada [7,8], els Boscos Aleatoris (Random Forests (RF)) [9] o, més recentment, l'aprenentatge profund (Deep Learning (DL)) [10] a diferents modalitats de neuroimatge ha ampliat les possibilitats d'anàlisi en imatges cerebrals molt més enllà de les tradicionals comparacions de grups cas-control.

Com a mostra d'en què s'està utilitzant els mètodes d'aprenentatge automàtic, molts investigadors han utilitzat dades basades en ressonància magnètica (RM) cerebral per detectar trastorns mentals. La majoria d'estudis s'han centrat a classificar controls sans i pacients [11–13]. Alguns altres estudis han fet servir imatges de ressonància magnètica per a discriminar el diagnòstic en diferents trastorns mentals [14–16]. En una revisió sistemàtica recent s'ha resumit la capacitat de predicció reportada en diferents estudis, concretament en classificar fent ús solament dades de RM cerebrals a pacients amb esquizofrènia respecte controls sans. D'entre 41 estudis, en 40 han reportat un rendiment dels models superior al 70% (percentatge d'encerts en el diagnòstic). D'aquests 40, en 35 s'havia fet servir RM funcional (fMRI) i, en 5, imatges de RM del tensor de difusió (DTI) [17]. Tot i que aquestes taxes de predicció poden semblar impressionants, la majoria d'aquestes eines de diagnòstic no s'han integrat encara a la pràctica clínica, ja que de moment la visita clínica ofereix un diagnòstic més acurat i fiable per l'especialista.

S'ha fet menys recerca en aprenentatge automàtic aplicat a l'estimació del risc de desenvolupar episodis futurs de psicosi (és a dir, recaigudes) [18,19]. Això és una llàstima perquè la detecció precoç podria endarrerir o fins i tot prevenir conseqüències futures greus [20]. Com a possibles variables predictores, alguns estudis han inclòs dades clíniques, com ara la presència de símptomes maníacs i negatius [21–23], el diagnòstic a l'inici juntament amb altra informació sociodemogràfica i d'escales clíniques [24,25], o el consum de drogues [24,26,27]. Altres han analitzat el poder predictiu de la informació biològica, com ara biomarcadors basats en la sang [28,29], dades genètiques [30–32], o la combinació de dades clíniques i biològiques [33].

Finalment, s'ha explorat encara menys l'ús de dades de RM per estimar quin pot ser el desenvolupament futur de la malaltia en un pacient, com poden ser els pacients en esquizofrènia. Només uns pocs estudis han pres aquest camí, utilitzant com a variables predictores els canvis en el volum cerebral durant el primer any després d'un primer episodi de psicosi (FEP) [34], anomalies anatòmiques cerebrals com ara el càvum septum pellucidum o hiperintensitats de matèria blanca [35,36], o dades basades en superfície cerebral o gruix cortical [36]. En tots aquests estudis han pres les variables, i han intentat estimar quina seria l'evolució del pacient, amb un encert d'entre el 60 i el 70%.

Tot i que els resultats encara són humils, aquests esforços de segur estan construint una pinya sobre la qual les futures investigacions aniran pujant pisos i crearan eines valuoses que podran ajudar al metge. Per exemple, aquestes eines podrien ajudar a detectar subjectes en risc de patir alguna malaltia o complicació o predir el grau de resposta a diferents tractaments amb l'objectiu final de millorar el benestar del pacient[37,38].

Finalment, fer aquestes estimacions a nivell individual permetria adaptar cada tractament a cada pacient, el que se sol anomenar medicina personalitzada o psiquiatria de precisió, i així sortir del paradigma de "one size fits all". Es podrien detectar, per exemple, grups de població amb necessitats especials per aconseguir una reducció més eficaç de símptomes, disfunció i recaigudes de pacients, aconseguint així una millora en la seva qualitat de vida. A més, això podria reduir el nombre d'hospitalitzacions relacionades amb les recaigudes (amb el consegüent estalvi econòmic pel sistema nacional de salut) i podria evitar la inclusió de pacients de baix risc en teràpies que són innecessàries per a ells (evitant, per tant, la iatrogènia farmacològica que malauradament encara és important en alguns contexts).

En els propers apartats d'aquesta introducció, explicaré alguns dels temes clau de la tesi. Primer explicaré breument el concepte de primers episodis psicòtics, ja que els pacients de l'estudi principal eren persones que havien patit un primer episodi psicòtic i l'objectiu era predir-ne el risc de recaiguda. A continuació introduiré en què consisteix la neuroimatge cerebral, perquè juntament amb certes dades clíniques, van ser les variables predictores utilitzades per estimar el risc de recaiguda. Seguidament, situaré en context els mètodes més usats en neuroimatge i com s'ubiquen dins el món de la intel·ligència artificial, un tema en boca de tots, però que molta gent desconeix. Per últim, dins d'aquesta introducció parlaré d'alguns punts crítics a evitar quan es treballa amb aquest tipus d'algorismes i dades, com a mostra de la cura i estima que hem tingut per a realitzar aquest treball.

1.1. PRIMERS EPISODIS PSICÒTICS

Un episodi psicòtic es defineix per la presència de símptomes com poden ser les idees delirants (ex., manté fermament creences infundades i considerades falses per la resta de persones de la seva subcultura), al·lucinacions (ex., la persona sent veus inexistents com a reals), llenguatge desorganitzat (ex., la seva parla és incoherent), i comportament catatònic o greument desorganitzat (ex., manté postures estranyes) durant un període de temps, generalment de setmanes o mesos. La psicosi és, per tant, una condició que afecta la forma en què el cervell processa la informació. Concretament, fa que es perdi el contacte amb la realitat, i es puguin percebre estímuls imaginaris com a reals. Els símptomes psicòtics solen classificar-se en positius o negatius. Els positius són aquells que s'afegeixen o distorsionen el funcionament normal de la persona, com poden ser els deliris, les al·lucinacions, o el pensament, llenguatge o comportament desorganitzat. Per contra, els símptomes negatius impliquen una reducció o pèrdua en el funcionament normal de la persona, com ara la falta d'interès o planificació, l'apatia, la manca de resposta emocional o l'aïllament social [39].

El primer episodi psicòtic sol venir acompanyat de la por, la confusió i l'angoixa, ja que el contingut dels deliris i les veus sovint és negatiu (ex., amenacen i insulten a la persona), i a més és una situació desconeguda per la persona que ho pateix. Això no només suposa un estressor per al mateix pacient, que pot viure l'episodi com a un fet traumàtic [40], sinó que sol implicar problemes mentals tals com l'angoixa en els seus familiars o cuidadors més propers [41]. Els primers episodis solen aparèixer sobretot durant les últimes etapes de l'adolescència o durant els 20-30 anys [42,43].

Després d'un primer episodi psicòtic (PEP) l'evolució pot variar des d'una remissió dels símptomes amb recuperació sostinguda, fins al desenvolupament de trastorns resistents al tractament antipsicòtic [44]. La psicosi es pot veure en diferents trastorns mentals, com pot ser l'esquizofrènia o altres condicions de l'espectre esquizofrènic, el trastorn bipolar, o el trastorn depressiu major [45].

1.1.1. POSSIBLES DIAGNÒSTICS DESPRÉS D'UN PRIMER EPISODI PSICÒTIC

Després d'un primer episodi psicòtic, aquest pot acabar derivant en futurs diagnòstics. Els possibles diagnòstics després d'un primer episodi psicòtic poden incloure:

- Trastorn psicòtic breu, trastorn esquizofreniform, i esquizofrènia: aquests trastorns es caracteritzen per la presència de símptomes psicòtics persistents (com ara al·lucinacions, deliris, pensament desorganitzat, discurs desorganitzat), que especialment en l'esquizofrènia també poden ser negatius (com ara aplanament afectiu o la pèrdua de motivació). Si hi ha símptomes afectius, no són clínicament significatius o són breus. Si el trastorn dura més d'un mes es parla de trastorn esquizofreniform, i si dura més de sis mesos, es parla d'esquizofrènia.

- Trastorn delirant: aquest trastorn es caracteritza per la presència de deliris persistents i no bizarres, sense altres símptomes psicòtics com al·lucinacions. Si hi ha símptomes afectius, no són clínicament significatius o són breus.

- Trastorns afectius (depressiu major o bipolar) amb símptomes psicòtics: quan també existeixen símptomes afectius clínicament significatius que compleixen totalment els criteris per un episodi d'alteració anímica.

- Trastorn esquizoafectiu: quan hi ha símptomes psicòtics i afectius i, contrari a l'esquizofrènia, els símptomes psicòtics són clínicament significatius i no breus, i contrari als trastorns afectius amb símptomes psicòtics, els símptomes psicòtics no apareixen només en el context dels episodis afectius.

- Trastorn psicòtic no especificat: aquest trastorn es dona quan els símptomes psicòtics no compleixen els criteris per a cap altre trastorn psicòtic específic.

- Trastorn psicòtic secundari a causa d'un altre trastorn mèdic o de substància: aquest trastorn es dona quan els símptomes psicòtics són causats per una condició mèdica o l'ús de substàncies.

1.1.2. IMPORTÀNCIA DE LA PREDICCIÓ DE LES RECAIGUDES

Si es parla de "primer" episodi psicòtic, és perquè és possible que no sigui l'únic. De fet, més de la meitat de persones que experimenten un primer episodi psicòtic n'experimentaran un altre abans de 3 anys [5,46] i fins a un 80% en els 5 anys vinents [5].

El desenvolupament posterior d'aquests primers episodis és molt important, ja que s'ha demostrat que les recaigudes durant els primers anys després del primer episodi són un factor important per predir l'evolució clínica i funcional a llarg-termini del pacient [47]. A més, cal tenir en compte que el fet de no tractar a temps la psicosi està associat amb una mala evolució tant dels símptomes negatius com dels positius, una reducció de les opcions que pugui remetre, i un empitjorament del funcionament social [48,49].

Per tal de poder considerar que hi ha una recaiguda en un pacient, primer ha d'haver-hi hagut un període de remissió sense símptomes. En aquesta tesi, seguint el criteri d'Andreasen et al. 2005, s'ha definit la remissió com l'absència de puntuacions superiors a 3 en els ítems de la *Positive and Negative Syndrome Scale* (PANSS)[50] P1, P2, P3, N1, N4, N6, PG5 i PG9 durant com a mínim 6 mesos; i recaiguda com la presència d'una puntuació superior a 3 en un d'aquests ítems durant com a mínim una setmana[51]. Aquests ítems corresponen a la presència de deliris (P1), desorganització conceptual (P2), conducta al·lucinatòria (P3), embotiment afectiu (N1), retraïment social (N4), falta d'espontaneïtat/fluïdesa a la conversa (N6), manierismes (moviments o postures artificials, PG5), i continguts inusuals del pensament (PG6).

1.2. LA RESSONÀNCIA MAGNÈTICA

La ressonància magnètica (RM) és una tècnica d'imatge mèdica no invasiva que utilitza ones de radio i un camp magnètic per produir imatges detallades dels teixits del cos. Utilitza propietats magnètiques dels teixits del cos per produir imatges que poden mostrar diferents tipus de teixits, com els ossos, els músculs, els tendons i els teixits del cervell. Aquestes imatges es poden utilitzar per diagnosticar i monitoritzar moltes condicions mèdiques, com ara tumors, malalties del cervell, malalties dels ossos i dels músculs, i malalties dels vasos sanguinis.

La RM funciona basant-se en els principis físics de la ressonància magnètica nuclear (RMN). El funcionament bàsic de la RM implica l'ús d'un gran camp magnètic per alinear els protons dels teixits del cos, especialment els protons d'hidrogen. Aquesta alineació es pot alterar aplicant una ona de radio de freqüència específica, que fa que els protons es desplacin fora de l'alineació. Quan els protons es tornen a alinear, emeten una senyal electromagnètica que es pot capturar i processar per produir imatges. La RM és capaç de produir imatges de diferents tipus de teixits del cos, ja que els teixits tenen diferents nivells de protons d'hidrogen i diferents temps de relaxació, el que fa que les imatges siguin diferents. Això permet que els metges puguin veure detalls precisos en una gran varietat de teixits i així ajudar en el diagnòstic de moltes malalties.

La informació provinent d'una imatge de ressonància magnètica es pot estructurar en vòxels. Un vòxel és una unitat tridimensional de dades en una imatge de volum, com ara una imatge de ressonància magnètica (MRI). Els vòxels són els elements bàsics d'una imatge tridimensional i són similars als píxels en una imatge bidimensional, de fet la paraula vòxel és una combinació de "volum" i de "píxel". Cada vòxel té una intensitat de senyal associada, que pot ser utilitzada per representar diferents característiques de l'estructura o la funció del cervell. Els vòxels es poden utilitzar per analitzar les dades de neuroimatge i permeten quantificar les diferencies en la intensitat de la senyal en diferents regions del cervell.

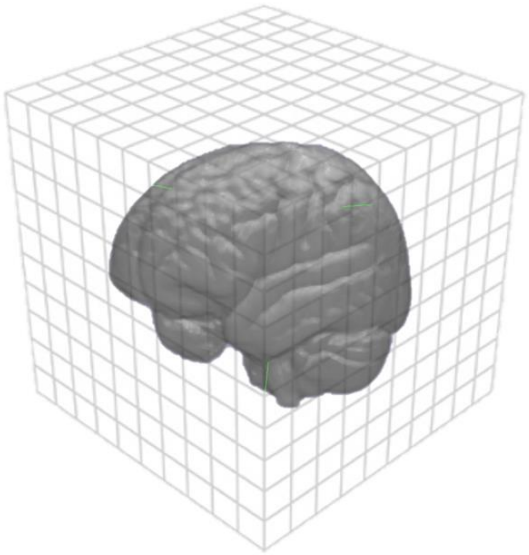


Figura 1. Concepte de vòxel. Cervell representat en vòxels.

1.3. NEUROIMATGE CEREBRAL

La neuroimatge és un terme genèric que es refereix a qualsevol metodologia que permet tenir una visualització estructural, funcional o farmacològica del sistema nerviós. En aquesta tesi em centraré en l'exploració cerebral mitjançant ressonància magnètica (RM). De forma general, podem classificar les tècniques de neuroimatge en dades de RM en dos tipus: estructural i funcional.

La RM cerebral estructural és una tècnica especialitzada en l'anàlisi i visualització de les característiques anatòmiques del cervell. Es pot utilitzar per exemple per a detectar dany cerebral o anomalies anatòmiques. En canvi, la RM funcional permet mesurar l'activitat cerebral detectant els canvis associats al flux de sang durant un període de temps.

Per aquesta tesi ens centrarem en les imatges de ressonància magnètica estructural. Concretament, plantejarem que les característiques anatòmiques del cervell podrien aportar informació rellevant sobre el desenvolupament futur d'una malaltia, en aquest cas el risc de recaure després d'un primer episodi psicòtic.

A continuació presentaré alguns dels principals passos de preprocessat estàndard que s'han de realitzar prèviament a utilitzar les imatges de ressonància magnètica estructural. En les

dades utilitzades a la tesi, aquest processat ha estat realitzat utilitzant el software Statistical Parametric Mapping 12 (SPM12) (https://www.fil.ion.ucl.ac.uk/spm/).

1.3.1. PREPROCESSAT DE LES IMATGES DE RESSONÀNCIA MAGNÈTICA ESTRUCTURAL

Inspecció visual (control de qualitat)

El primer pas, abans de prosseguir amb altres etapes del processament, consisteix a fer un control de qualitat de les imatges crues obtingudes de l'escàner. Així podrem detectar imatges que serien inútils pels nostres interessos posteriors. Cal comprovar que les imatges s'hagin guardat bé (a vegades només es guarda una part de la imatge), detectar artefactes de moviment (deguts al fet que la persona mogui excessivament el cap durant l'adquisició de la imatge), o comprovar que les imatges estan correctament orientades (ex., el lòbul frontal a davant i l'occipital a darrere).

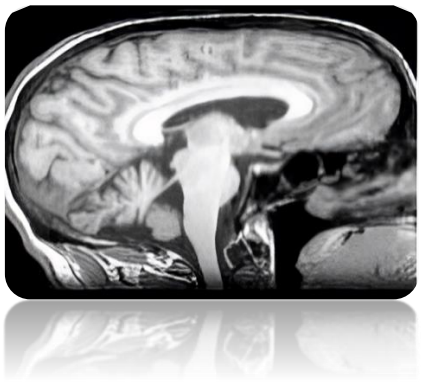


Figura 2. Imatge de RM cerebral previ al preprocessat.

Segmentació

La segmentació consisteix a classificar els vòxels cerebrals en diferents grups que representen diversos tipus de teixit: substància grisa, substància blanca, líquid cefalorraquidi (LCR), i altres no cerebrals (crani, tendons/teixit tou i aire/fons). Aquesta classificació es fa mirant el senyal de cada vòxel en la seqüència T1 original. El procés pot incorporar alguns factors de correcció que milloren el resultat final, com pot ser tenir en compte la seva localització o els vòxels veïns per tal d'assignar un tipus de teixit o un altre. Com que els diferents teixits tenen uns valors

d'intensitat diferents, al final obtindrem sis imatges corresponents a cadascun dels teixits cerebrals (substància grisa, blanca, i LCR) i als no cerebrals (crani, tendons/teixit tou i aire/fons).

És important esmentar que en la majoria de casos els algoritmes de segmentació no classifiquen cada vòxel de manera excloent en una sola categoria de teixit cerebral. De fet, els algoritmes de segmentació assignen una probabilitat a cada vòxel de pertànyer a cadascun dels teixits, per la qual cosa cada vòxel (i, per tant, cada regió cerebral) estarà representat mitjançant un valor probabilístic en les imatges corresponents als diferents teixits cerebrals.

Figura 3. Exemples d'imatges cerebrals segmentades: imatge de matèria grisa (superior) i de matèria blanca (inferior).

Normalització espacial

La forma i la mida global dels cervells de les persones són diferents, per això, per fer comparacions estadístiques entre grups de subjectes, s'han d'eliminar aquestes diferències abans de l'anàlisi estadística. Els algoritmes de normalització ho fan deformant cada cervell per portar-los a un espai estereotàctic comú, així les diferències en forma i mida global desapareixen, però es mantenen les diferències interindividuals de cada teixit cerebral.

Els atles més usats per realitzar aquesta normalització són l'atles de Talairach, basat en un sol cas, i l'atles de l'Institut Neurològic de Mont-real (MNI) basat en 305 cervells. La normalització permet generalitzar la localització anatòmica dels resultats, així només cal reportar les coordenades x, y i z i l'atles utilitzat per saber la zona cerebral de la qual es parla.

Modulació

El procés de normalització, tanmateix, té un efecte no desitjat sobre la interpretació de les dades. Quan es deformen les imatges a un espai comú, si bé es mantenen els valors de probabilitat per a cadascun dels teixits cerebrals derivats de la segmentació, es perd informació sobre les diferències volumètriques regionals dels individus. Com que aquesta informació pot ser rellevant, es pot afegir al processament de dades un apartat de modulació, que consisteix a multiplicar cada vòxel per un valor que representa la deformació que ha patit el vòxel en qüestió durant la normalització. Així les regions que s'han expandit en la normalització veuen el seu valor reduït, i els vòxels que s'han empetitit veuen el seu valor augmentat. D'aquesta forma reincorporem la informació volumètrica perduda durant la normalització espacial, i en aquest moment, el valor ja no només indica la probabilitat de pertànyer a un teixit o un altre, sinó el volum original del teixit (ex., ml de substància grisa).

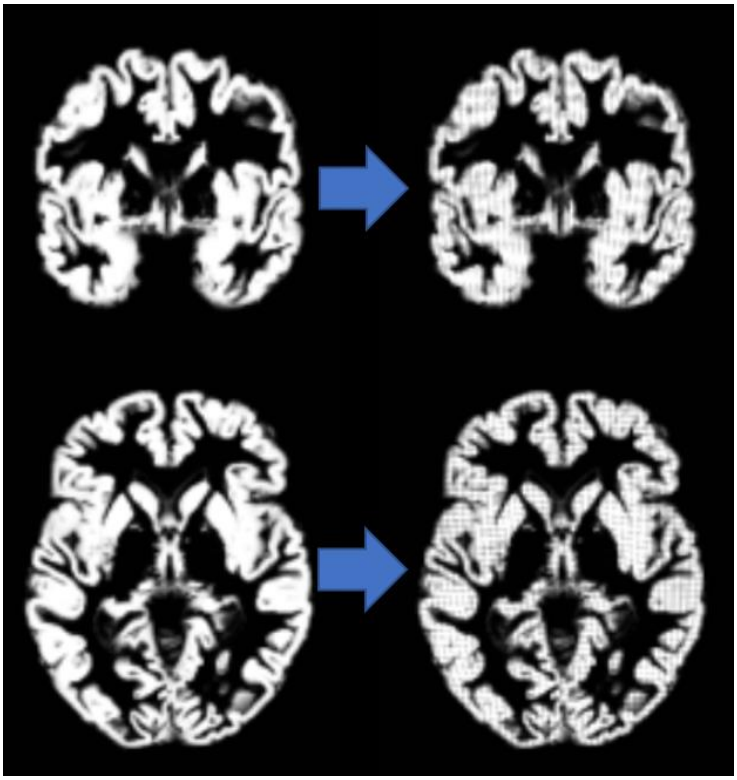


Figura 4. Imatge sense modular (esquerra) i imatge un cop aplicada la modulació (dreta).

Suavitzat

Després de la segmentació, normalització i modulació, les imatges són finalment suavitzades. Aquest suavitzat (o smoothing) s'aconsegueix fent una mitjana ponderada del valor de cada vòxel amb els dels seus vòxels més pròxims. Així obtenim uns valors més coherents al llarg de l'estructura en qüestió, i a més desapareixen els valors que podrien ser soroll o artefactes i només apareixen en un sol vòxel. Això a més implica que el valor del vòxel ja no solament representa aquest vòxel de forma singular, sinó que incorpora informació dels vòxels veïns i, per tant, és representatiu del volum d'una esfera al voltant del vòxel inicial. Després d'aplicar el suavitzat, les imatges podran ser analitzades estadísticament. Normalment, per fer el suavitzat es fa servir un kernel Gaussià amb una amplitud màxima en el punt central (o en anglès full width at half maximum -FWHM-) d'entre 4 i 12 mm. Per consegüent, aquest suavitzat contribueix a fer que les dades tinguin una distribució normal de les dades, cosa que facilita la seva anàlisi estadística.

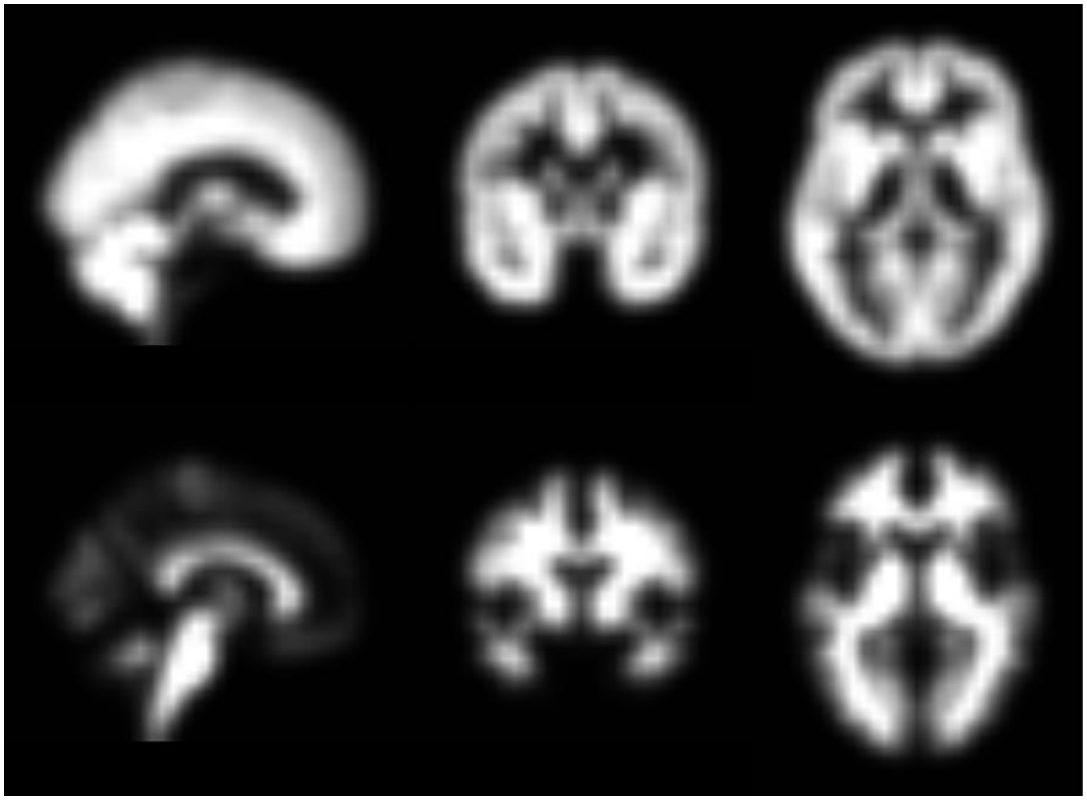


Figura 5. Imatges amb tot el preprocessat aplicat de matèria grisa (superior) i matèria blanca (inferior).

1.4. APRENENTATGE AUTOMÀTIC

L'Aprenentatge Automàtic (AA), o Machine Learning en anglès, és un dels principals subcamps de l'especialitat anomenada Intel·ligència Artificial (IA). Abans d'entrar en termes més concrets, ubiquem breument com es relacionen.

La **Intel·ligència Artificial** és un terme que fa referència a la capacitat d'un sistema a interpretar dades externes, aprendre d'aquestes dades, i posteriorment utilitzar aquest aprenentatge per aconseguir un objectiu final, ja sigui predir un valor, classificar en categories o interpretar dades no estructurades.

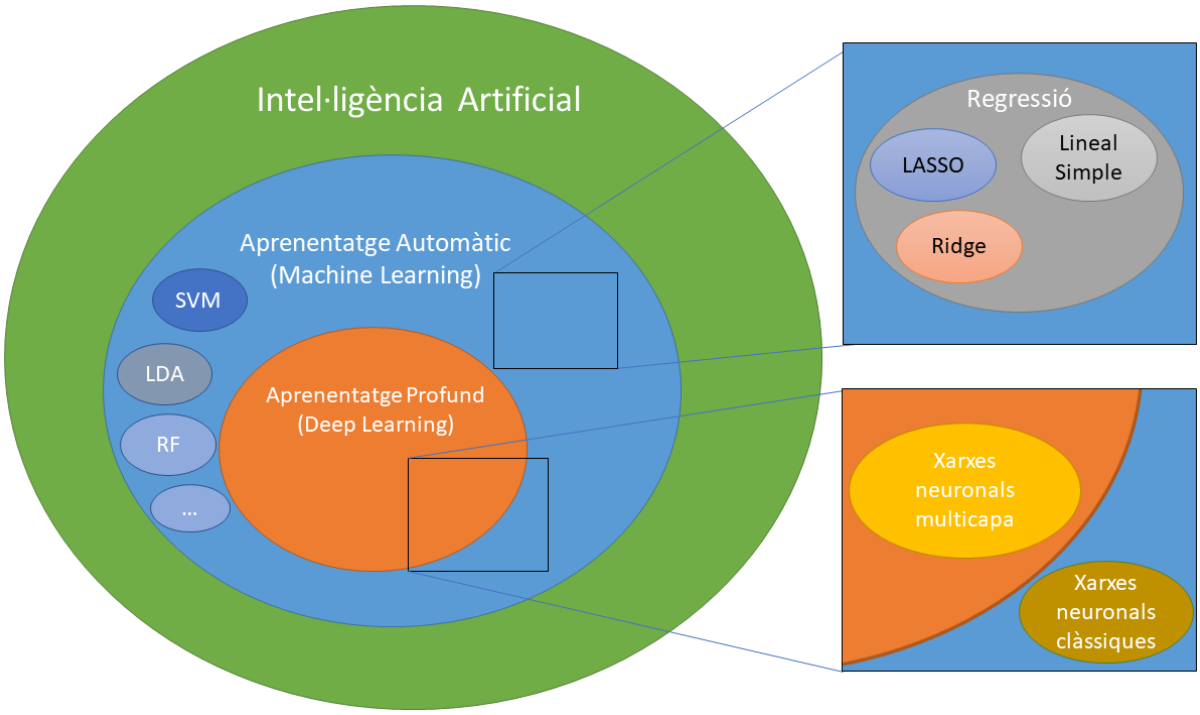


Figura 6. Esquema de la família de la Intel·ligència Artificial. Representació amb alguns dels principals algoritmes en relació a la IA.

Hi ha tres principals categories d'aprenentatge automàtic: l'aprenentatge supervisat, l'aprenentatge no supervisat i l'aprenentatge per reforç.

En **l'aprenentatge supervisat** el model aprèn basant-se en la informació que li passem. És com si a un nen petit li volem ensenyar a distingir gossos de gats. Primer li ensenyarem un gos i li direm que allò és un gos. Seguidament, també li ensenyarem un gat i li direm que allò

que veu és un gat. El nen (que en aquesta analogia faria el paper de model, ja que ell aprèn partint d'unes dades, igual que un model) acabaria podent dir-nos si aquell animal que veu és un gos o un gat.

En **l'aprenentatge no supervisat** el model intenta entendre les dades pel seu compte amb una intervenció mínima de la persona que dissenya l'algorisme. Seguint amb l'exemple dels gats i els gossos, aquí seria com si li ensenyéssim diferents animals al nen i ell entengués pel seu compte que hi ha dos tipus d'animals diferents i creés les seves normes internes per a diferenciar-los. Al final el nen (que ja he dit que fa el paper de model) ens podria dir que ell veu dos tipus diferents d'animals i ens diria quins són quins, tot i que no sabria dir-nos quin nom té cadascun.

En **l'aprenentatge per reforç** tal com diu el nom el model rep informació només sobre si el que està fent és correcte o incorrecte. Per enèsima vegada, seguint amb el nen i els animals, aquí el procés seria que el nen, mirant els animals, anés dient "això és un gos", "això és un gat", i un agent li aniria dient si aquell seguit de decisions i afirmacions que va fent estan bé o no, i li donaria un "premi". El nen en aquest cas intentaria optimitzar els beneficis d'aquest premi, anant reajustant les característiques del model que li permeten aconseguir una resposta per tal de maximitzar els beneficis finals.

A continuació en veurem alguns dels principals, centrant-nos en la seva aplicabilitat a la neuroimatge en psiquiatria, i més concretament a trastorns relacionats amb els primers episodis psicòtics i l'esquizofrènia.

1.5. APRENENTATGE AUTOMÀTIC EN NEUROIMATGE

El camp de l'aprenentatge automàtic engloba una extensa llista d'algorismes, cadascun amb els seus punts forts i febles. Cadascun d'aquests algorismes s'hauria d'ajustar per obtenir una compensació entre el model que no s'ajusta prou bé a les dades, el que s'anomena infraajustament (underfitting), i l'ajust excessiu de les dades d'entrenament, anomenat sobreajustament (overfitting). Un model poc adaptat no aconsegueix capturar la relació entre

les dades de ressonància magnètica i la resposta, per la qual cosa funciona malament fins i tot en les dades d'entrenament. El sobreajustament es produeix quan el model troba relacions entre les dades de la ressonància magnètica i el resultat que només es basen en detalls aleatoris i particulars de les dades d'entrenament i, per tant, el seu rendiment és pobre en noves dades. Sobre la gran selecció d'algorismes, revisarem només els més comuns en neuroimatge.

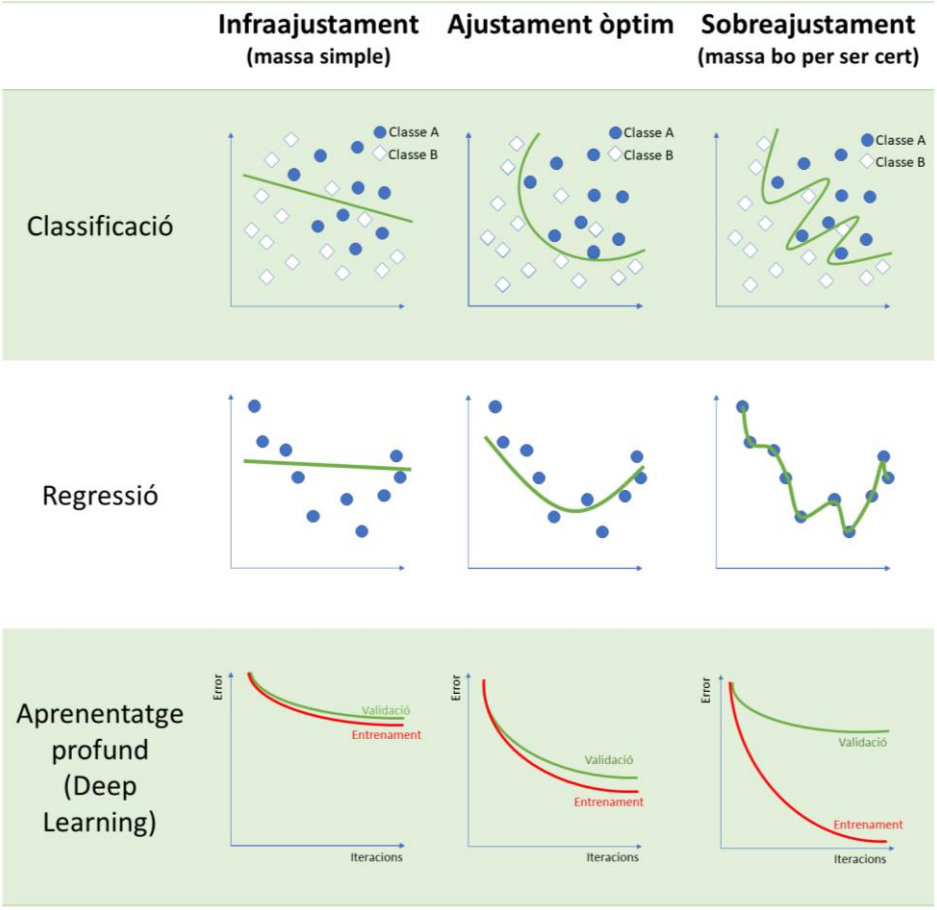


Figura 7. Els diferents tipus d'ajustaments dels models a les dades. Infraajustament, ajustament òptim i sobreajustament.

1.5.1. REGRESSIÓ DE LASSO

La regressió de Lasso (acrònim de Least Absolute Shrinkage and Selection Operator) és un tipus de regressió regularitzada que en els seus models tendeix a contreure els coeficients d'aquelles variables que no contribueixen gaire en la predicció, fins i tot fent que es descartin completament algunes variables en el model final [52]. Això implica que els models que s'acaben

obtenint solen ser senzills i amb menys paràmetres que els que obtindríem amb una regressió lineal normal per exemple. Aquest tipus de model, és especialment útil per dades que mostren alts nivells de multicol·linealitat o també en processos on ens interessa poder seleccionar aquelles variables més útils (un procés anomenat selecció de variables o feature selection en anglès).

Tot i ser un algorisme senzill, ha donat resultats positius per detectar subjectes amb risc molt alt de desenvolupar episodis psicòtics o esquizofrènia i fins i tot ha funcionat millor que algorismes més complexos [53]. En una comparativa entre models en predicció utilitzant dades de neuroimatge, Salvador et al. van avaluar el poder discriminador d'alguns dels algorismes més utilitzats en l'ISR per a la predicció en psicosi. Van comparar sistemàticament els diferents algorismes i, entre ells, els classificadors Ridge, LASSO i Elastic Net van tenir un rendiment similar, si no lleugerament millor, que altres classificadors [54].

1.5.2. SUPPORT VECTOR MACHINES

La màquina de vector de suport (SVM) és un mètode de classificació discriminatori supervisat que crea hiperplans que fan de frontera i permeten separar òptimament les dades en diferents grups, és a dir que intenta que la distància entre les dades d'entrada i l'hiperplà frontera siguin tan grans com sigui possible. Es poden utilitzar tant per a tasques de classificació com de regressió. Un punt fort de l'algorisme és la capacitat de realitzar classificacions no lineals. Això es fa convertint les dades de l'espai d'entrada, que tenien una dimensió baixa, i es converteixen a un espai de dimensió superior per tal que l'hiperplà pugui separar les classes de forma òptima. Per millorar la generalització del model i evitar el sobreajustament, s'ha de considerar una compensació entre maximitzar el marge entre les classes i minimitzar el nombre de classificacions errònies.

Els algorismes SVM s'han establert com a poderoses eines per a trobar biomarcadors neuroanatòmics objectius, ja que poden manejar dades d'alta dimensionalitat de manera molt eficaç, considerar correlacions interregionals entre diferents regions cerebrals i fer inferències a un nivell d'un sol subjecte amb un resultat de classificació decent [55].

Probablement, és un dels algorismes de classificació d'aprenentatge automàtic més utilitzats en dades de neuroimatge. Un enfocament comú ha estat l'ús de SVM per a classificar els subjectes amb risc ultra alt de psicosi respecte de controls sans [56–58]. A causa de la seva capacitat per a tractar dades d'alta dimensionalitat, també s'ha aplicat per a classificar entre depressió d'inici recent i psicosi d'inici recent fent servir tant informació neuroanatòmica com dades clíniques [59], per a trobar subtipus neurocognitius basats en el rendiment cognitiu i alteracions neurocognitives en la psicosi d'aparició recent [60,61], per identificar pacients amb esquizofrènia en funció de regions subcorticals [62] o dades de connectivitat funcional de xarxes [63]. Es va provar un enfocament multimodal que combinava ressonància magnètica estructural, imatges de tensors de difusió i dades de ressonància magnètica funcional en estat de repòs per classificar pacients amb esquizofrènia crònica vs. pacients amb PEP comparant diferents algorismes com Random Forest (RF), regressió logística, Linear Discriminant Analysis (LDA), K-Nearest Neighbor classification (KNN) i SVM, resultant aquest últim com el model amb millor rendiment [64]. Steardo et al. en una revisió sistemàtica recent, van analitzar 22 estudis que utilitzaven SVM sobre RM funcional fent servir diferents biomarcadors per classificar entre pacients amb esquizofrènia i controls sans, on 19 dels 22 estudis van reportar una prometedora precisió >70% [65].

La popularitat de les SVM i la compensació entre rendiment i simplicitat són els seus principals punts forts. És fonamental tenir en compte que el seu rendiment depèn en gran manera dels hiperparàmetres escollits, i aquests solen ajustar-se mitjançant un mètode de cerca en quadrícula (grid search) [66].

1.5.3. LINEAR DISCRIMINANT ANALYSIS

De manera similar a SVM, l'anàlisi discriminant lineal (LDA) implica transformar la dimensionalitat de les dades. Però contràriament a SVM, en els LDA, les dades es projecten en un espai de dimensions inferiors on els diferents grups de dades es poden separar òptimament utilitzant un nucli (o kernel) no lineal [67].

S'ha utilitzat en tasques de classificació com l'esquizofrènia d'inici recent (ROS) vs. HC [68], pacients amb PEP vs. HC [69–71] o pacients amb esquizofrènia (SZ) vs. HC [72–75] amb precisions

superiors al 70%. Winterburn et al. van utilitzar tres conjunts de dades independents per validar el poder discriminador de l'LDA per classificar els pacients amb SZ de HC, utilitzant diferents dades de neuroimatge. Van comparar dades de grossor cortical, mapes de corbes i VBM modulades, i van resultar que, utilitzant el seu conjunt de dades més gran, la precisió va ser lleugerament inferior en comparació amb articles anteriors, on hi havia conjunts de dades més petits [76].

La principal fortalesa de l'LDA és que no és tan propens a patir sobreajustament i que sol tenir cost computacional baix, mitjançant la reducció de l'espai dimensional. Malgrat això, el seu principal inconvenient és que requereix assumir que la matriu de covariància en els grups de dades és idèntica, cosa poc freqüent en dades del món real.

1.5.4. ARBRES DE DECISIÓ I BOSC ALEATORI

Els arbres de decisió són mètodes d'aprenentatge supervisat no paramètrics que s'utilitzen tant per a la classificació com per a la regressió. Poden predir valors mitjançant l'aprenentatge de regles de decisió senzilles inferides a partir de les característiques de les dades. Aquests algorismes tendeixen a sobreajustar-se i, per tant, a no generalitzar-se bé a noves dades. Per superar aquesta limitació, existeix una variació sobre aquest algorisme anomenada Bosc Aleatori (Random forest, RF). Es tracta simplement d'una col·lecció d'arbres de decisió els resultats dels quals s'agreguen en un únic resultat al final. Els RF incorporen interaccions entre predictors en el model, detectant relacions tant lineals com no lineals.

S'ha utilitzat per a classificar grups, com ara pacients amb esquizofrènia d'inici infantil i controls sans [77] o esquizofrènia, trastorn bipolar i controls sans [78,79].

És un algorisme que generalment proporciona una alta precisió i un equilibri entre la compensació biaix-variància. Els seus principals inconvenients són que tendeix a ser computacionalment intens en grans conjunts de dades. A més, pot ser difícil interpretar-ne els resultats, ja que és difícil analitzar-ne tots els coeficients.

1.5.5. XARXES NEURONALS I APRENENTATGE PROFUND

Les Xarxes Neuronals Artificials (ANN) són una família d'algorismes d'aprenentatge automàtic inspirats en el funcionament biològic del cervell. De forma aproximada al que fa el nostre cervell, aquests algoritmes tenen neurones que reben un senyal, el processen, envien un senyal a la següent neurona connectada, etc. fins a obtenir un resultat final. Per ajustar les capacitats d'aprenentatge del model, cada neurona i sinapsi pot tenir pesos per augmentar o disminuir la força del senyal. Les neurones s'agreguen en capes i, quan augmenta el nombre de capes, l'algorisme es coneix com a Deep Learning. Aquests models avançats poden extreure característiques latents complexes de dades originals mínimament preprocessades mitjançant transformacions no lineals. Per evitar el sobreajustament, existeix un mètode anomenat Dropout. El Dropout és un mètode de regularització que ignora o "deixa caure" aleatòriament alguns nodes de la capa, cosa que és similar a afegir soroll al procés d'entrenament. Això millora la generalització del model i redueix el sobreajustament.

L'aprenentatge profund s'utilitza sovint per classificar, predir valors o fins i tot detectar o segmentar regions del cervell.

L'aprenentatge profund (o Deep Learning) és un camp molt ampli, que ha augmentat el rendiment en alguns problemes de classificació/predicció a causa de trobar patrons complexos en dades altament complexes. Tot i ser àmpliament utilitzat per a detectar automàticament tumors [80] o detectar lesions d'esclerosi múltiple [81] en imatges d'RM cerebral, encara no s'utilitza àmpliament en la detecció de trastorns de salut mental o en l'estimació de riscos d'una futura evolució de la malaltia en salut mental.

Un dels problemes que té l'aprenentatge profund sol ser la interpretabilitat dels models, o el que se sol anomenar el problema de la caixa negra. Per a superar aquesta problemàtica en xarxes neuronals artificials, s'estan desenvolupant diferents enfocaments, com la creació de mapes de calor utilitzant Layer-Wise Propagation per a identificar les característiques més importants implicades en les decisions que pren l'algorisme [82,83].

1.5.6. ANÀLISI DE SUPERVIVÈNCIA

L'anàlisi de supervivència és un conjunt de tècniques que permeten estudiar el temps que passa fins a un esdeveniment. En la recerca mèdica, originàriament, aquest esdeveniment solia ser la mort d'un subjecte, i d'aquí n'hi ve el nom. Però a la pràctica, aquest tipus d'anàlisi es pot aplicar a qualsevol esdeveniment, tal com pot ser l'objectiu d'aquesta tesi, la recaiguda d'un pacient que havia remès els símptomes d'un primer episodi psicòtic. Aquest tipus d'estudi és especialment útil en situacions en què hi poden haver censures, per exemple quan tenim subjectes que l'únic que sabem és que passat un temps no ha desenvolupat l'esdeveniment estudiat (ja sigui perquè hi ha hagut una pèrdua de seguiment o perquè s'ha arribat al final de l'estudi i el pacient no ha recaigut).

Un model d'especial interès dins de l'anàlisi de supervivència, i que és el que hem utilitzat en l'article principal de la tesi, és el model de regressió de Cox, altrament conegut com a model de riscos proporcionals. És un model que permet tenir en compte diferents variables a la vegada i que analitza la relació entre la distribució de supervivència i aquestes variables [84]. Ha estat utilitzat, per exemple, per investigar si els trastorns del son prediuen la conversió a psicosi en els dos anys vinents des que es diagnostica el trastorn [85].

1.6. ERRORS COMUNS

Ara que hem vist els principals fonaments de la tesi, permeteu-me que expliqui alguns dels punts especialment importants que hem tingut en compte a l'hora de realitzar l'article principal, i que permeten entendre millor el procés en la presa de decisions metodològiques de l'estudi.

1.6.1. REPRODUCTIBILITAT I REPLICABILITAT

La reproductibilitat (reproducibility en anglès) usualment es refereix al fet d'obtenir (aproximadament) els mateixos resultats d'un article si utilitzem els mateixos procediments que en l'article sobre les mateixes dades. En canvi, la replicabilitat (replicability) sol implicar

que usant els mateixos procediments que en el mètode original, es puguin aconseguir resultats consistents sobre noves dades.

En els estudis de neuroimatge normalment s'han de prendre moltes decisions durant els processos de control de qualitat de les imatges, el preprocessat o les anàlisis estadístiques. A més, en els mètodes d'aprenentatge automàtic, cada algorisme té diferents paràmetres que els autors han de definir per tal de trobar la millor aproximació a les seves anàlisis.

Per aconseguir la reproductibilitat, totes les decisions preses amb relació al processament de les imatges i els paràmetres dels models han de ser detallades i compartides pels autors. Idealment, els articles haurien de reportar totes aquestes dades. Addicionalment, fer públics els models, codi i dades hauria de ser una pràctica comuna per tal de permetre a terceres parts examinar tot el procés de l'anàlisi a fons i així facilitar reproduir tot el procés que s'hagi aplicat en un estudi. Tot i això, les dades, especialment quan parlem de dades mèdiques, solen no poder-se compartir per motius de privacitat.

En canvi, per aconseguir la replicabilitat, s'hauria d'encoratjar la realització d'estudis independents de replicació. Desafortunadament, una limitació comuna que prevén l'aplicabilitat de molts algorismes reportats en estudis és la seva baixa replicabilitat, la baixa generalització dels mètodes. Molts models solen ser desenvolupats en mostres petites o en dades provinents d'un sol centre a causa de la dificultat tant en temps com en diners d'obtenir mostres més grans. De fet, per molt que aquests models fets amb mostres petites o d'un sol centre tinguin un bon rendiment sobre la mateixa mostra, el sobreajustament al que ha estat exposat el model fa que aquest rendiment baixi quan s'apliquen les tècniques a noves dades. Com en l'aprenentatge humà, com més exemples vegi un algorisme, més serà capaç d'extrapolar el seu aprenentatge a noves dades [86]. Per exemple, recentment, alguns autors han provat d'avaluar alguns models publicats utilitzant dades neuropsicològiques i clíniques per tal de predir la transició a psicosi en diferents subjectes en alt risc clínic de psicosi. Quan aquests models van ser aplicats sobre mostres noves, els models que havien estat publicats amb alts índex d'èxit, sobre les noves dades han estat incapaços de predir o han mostrat nivells insuficients de predicció [87,88].

No és infreqüent en l'aprenentatge automàtic de fer una bateria de tests classificadors o regressors i acabar reportant aquell que ha tingut un millor rendiment. Tanmateix, la ciència

hauria d'intentar evitar basar els resultats en l'anomenada "fal·làcia de prova incompleta" o "cherry Picking" en anglès, és a dir evitar escollir només aquells resultats profitosos pel mateix estudi. Hi poden haver excepcions i en alguns casos pot ser la millor forma de procedir, però en general això pot portar fàcilment a un sobreanàlisi de les dades.

1.6.2. BIAIX

Qualsevol model o representació de dades és susceptible de tenir algun biaix, ja que aconseguir. En crear o validar models d'aprenentatge automàtic hi ha diferents passos que poden ser fonts de biaix [89].

Per exemple, molts estudis utilitzen un enfocament dividit en dues parts. La primera part consisteix a crear un model (per exemple seleccionar quines variables prediuen millor una resposta). La segona part sol ser la validació d'aquest model (com pot ser aplicar el model per estimar com de bé funciona el model creat). Desgraciadament, si els investigadors fan servir les mateixes dades per crear el model i validar-lo, el rendiment estimat estarà inflat [90]. És per això que per a fer la validació i la creació del model (fins i tot en l'etapa de la selecció de variables) sol ser molt recomanable fer servir unes dades diferents.

Un altre exemple, important per a aquesta tesi, és l'ús de dades multicèntriques. Donada la dificultat d'obtenir grans bases de dades mèdiques, és habitual la col·laboració entre centres. Cada centre pot tenir certs biaixos en les seves dades, en el cas de ressonàncies magnètiques la mateixa màquina pot tenir diferències entre un centre i un altre, entre un model de màquina i una altra, les actualitzacions de programari que tinguin, els paràmetres... És per això que si no es tenen en compte aquestes diferències entre centres, els models d'aprenentatge automàtic podrien utilitzar de forma "fraudulenta" les diferències que troba entre centres per a predir una resposta. Aquests potencials efectes del centre s'han de controlar cautament, ja que ignorar-los podria produir un rendiment inflat, fins i tot encara que els models no prediguin de forma correcta [91,92].

Una forma de controlar els efectes derivats del centre en neuroimatge és utilitzar alguna tècnica d'harmonització de les dades. Un dels mètodes que es poden fer servir, i el qual ha

estat utilitzat en aquesta tesi, és el ComBat[1]. Tot i no formar part de l'estructura formal de la tesi, com a coautor l'explicaré de forma breu a l'apartat de mètodes per tal d'aclarir com hem gestionat el fet de tenir dades de diferents centres.

1.6.3. UTILITAT CLÍNICA

Una preocupació comuna entre professionals clínics és la dubtosa utilitat clínica de molts models d'aprenentatge automàtic recentment desenvolupats [93,94]. D'una banda, l'aprenentatge automàtic en dades mèdiques ha demostrat ser una eina impressionant per replicar i automatitzar processos humans, com ara la detecció automàtica per ordinador (CAD) de lesions en exploracions cerebrals, exploracions corporals o mamografies [95]. No obstant això, estudis consistents en detectar si una imatge d'RM pertany a un pacient o a un control sa poden tenir una utilitat clínica limitada [76]. Si bé, s'ha de reconèixer que aquests estudis són realment valuosos com a prova de concepte, hem d'assegurar-nos progressivament que els metges puguin trobar-los útils, és a dir, que la pregunta a la qual dona resposta el model estigui en línia amb les necessitats clíniques.

En aquest sentit, és essencial mantenir una distinció entre el que és un "model" i el que és una "eina". Un model pot ser necessari per a una investigació posterior o amb finalitats metodològiques. Per contra, una eina hauria de ser útil, factible i segura per a la presa de decisions clíniques en entorns del món real [94].

[1] L'article de l'adaptació del mètode per a dades de neuroimatge estructural es pot trobar com a article adjunt a aquesta tesi.

2. OBJECTIUS

L'objectiu general d'aquesta tesi va ser la creació i validació d'un model d'estimació del risc de recaiguda després d'un primer episodi psicòtic a partir de dades clíniques i de ressonància magnètica cerebral estructural, usant dades de diferents centres de l'estat.

Abans d'abordar aquest objectiu, però, vam decidir resoldre dues qüestions metodològiques importants: a) vam voler analitzar si, quan es crea un model de predicció usant dades multicèntriques, el fet de combinar dades de diferents centres pot esbiaixar l'estimació de la precisió del model, encara que s'hagin usat mètodes per eliminar les diferències entre centres; i b) vam voler determinar quines són les característiques de la ressonància magnètica cerebral estructural i els paràmetres de les anàlisis que poden optimitzar la precisió d'un model de predicció.

Per aquests motius, els principals objectius d'aquesta tesi han estat els següents:

1. Comprovar si, quan es crea un model de predicció usant dades multicèntriques, existeixen efectes del centre que poden esbiaixar l'estimació de la precisió del model, encara que s'hagin usat mètodes per eliminar les diferències entre centres. [article 1] [92]

2. Si fos aquest el cas, trobar mètodes perquè puguin evitar aquest biaix. [article 1][92]

3. Trobar les característiques de la ressonància magnètica cerebral estructural i els paràmetres de les anàlisis que millorin la precisió d'un model de predicció. [article 2][96]

4. Crear i validar un model de predicció per estimar el risc de recaiguda després d'un primer episodi psicòtic a partir de les dades de ressonància magnètica cerebral estructural i les dades clíniques. [article 2] [96]

5. Observar si la combinació de dades de ressonància magnètica cerebral estructural i dades clíniques augmenta la precisió de les prediccions, en comparació a models

basats només en dades de ressonància magnètica o basats només en dades clíniques. [article 2] [96]

A més a més, també teníem com a objectius proporcionar obertament a la comunitat científica:

6. Un software per estimar de forma fàcil la precisió d'un model de predicció amb dades multicèntriques evitant el possible biaix relacionat amb els efectes del centre. [article 1] [92]

7. Un software per crear de forma fàcil un model de predicció per estimar el risc de recaiguda (o altres variables d'interès clínic) a partir de les dades de ressonància magnètica cerebral estructural i les dades clíniques amb les possibles optimitzacions trobades durant la tesi. [article 2] [96]

8. El model de predicció per estimar el risc de recaiguda després d'un primer episodi psicòtic creat a aquesta tesi, per tal de facilitar-ne la validació externa, és a dir, que pugui ser validat per tercers en noves mostres de pacients. [article 1] [92]

3. HIPÒTESIS

Les hipòtesis principals que hem seguit són:

1. Quan es creen models de predicció amb dades de diferents centres, existeixen efectes del centre que esbiaixen l'estimació de la precisió del model, encara que s'hagin usat mètodes per eliminar les diferències entre centres. [article 1][92]

2. Existeixen mètodes que eviten aquest biaix. [article 1][92]

3. Durant la creació d'un model de predicció basant en dades de ressonància magnètica cerebral estructural, existeixen característiques de les dades i paràmetres de les anàlisis que milloren la precisió del model. [article 2][96]

4. Es pot crear un model de predicció que estimi exitosament el risc de recaiguda després d'un primer episodi psicòtic a partir de les dades de ressonància magnètica cerebral estructural i les dades clíniques [article 2][96]

5. La precisió de les prediccions és major en aquest model, que combina dades de ressonància magnètica cerebral estructural i dades clíniques augmenta la precisió de les prediccions, que en els models basats només en dades de ressonància magnètica o basats només en dades clíniques.

Els objectius 5-7 no tenen hipòtesis associades ja que consistien en proporcionar resultats del treball a la comunitat científica.

4. MÈTODES I RESULTATS: RESUM I DISCUSSIÓ GLOBAL DELS RESULTATS

Aquesta tesi està formada formalment per un article adjunt del que n'he sigut coautor, sobre els mètodes per eliminar els efectes del centre en estudis de ressonància magnètica multicèntrics, i que per tant guarda una forta relació amb els articles principals i l'enfocament general de la tesi:

- Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage **218**, (2020) [91]

I per dos articles principals:

- Article 1: Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. *Psychiatry Res. Neuroimaging* **314**, 111313 (2021) [92]

- Article 2: Combining MRI and clinical data to detect high relapse risk after the first episode of psychosis. *Schizophr. Heidelb. Ger.* **8**, 100 (2022) [96]

A continuació en faré un breu resum de cadascun dels tres. Els dos articles de la tesi estan adjunts estan adjunts a la secció 5. L'article adjunt es troba a la secció Material adjunt al final de la tesi.

4.1. ARTICLE ADJUNT: INCREASED POWER BY HARMONIZING STRUCTURAL MRI SITE DIFFERENCES WITH THE COMBAT BATCH ADJUSTMENT METHOD IN ENIGMA

4.1.1. INTRODUCCIÓ

És comú, que en els estudis de neuroimatge s'utilitzin dades de diferents centres per intentar solucionar la problemàtica que té obtenir mostres grans. Una de les opcions que han sortit darrerament per tal de facilitar aquesta col·laboració entre centres són els consorcis, com ara el "Enhancing Neuro Imaging Genetics through Meta-Analysis" (ENIGMA) Consortium. No obstant això, aquesta compartició de moltes dades provinents de molts centres diferents implica un augment de l'heterogeneïtat deguda a les diferències dels dispositius i les seqüències emprades.

En l'estudi, vàrem comparar el mètode ComBat-Mega respecte a l'ús de dues altres tècniques molt utilitzades pel mateix fi, el metaanàlisis d'efectes mixtos (RE-Meta) i utilitzar el centre com a un factor aleatori dins un model lineal d'efectes mixtos (ME-Mega).

L'ús de dades agregades per centre en l'estimació i control de l'heterogeneïtat pot ser subòptim en el RE-Meta [97]. I en el ME-Mega, com també al RE-Meta s'assumeix que l'error segueix una distribució normal igual en tots els centres, cosa que no acostuma a passar, ja que solen tenir diferents errors de variàncies. A més, aquests dos mètodes estimen l'heterogeneïtat de les diferents regions d'interès de forma separada, mentre que és probable que totes les regions comparteixin algun nivell d'heterogeneïtat. Per intentar millorar les limitacions dels mètodes esmentats vam provar l'ús del ComBat [98], un mètode utilitzat originàriament en dades genòmiques i que ja havia estat provat amb èxit en mostres petites de centres (<7 centres) [99].

En aquest article, hem provat l'ús del mètode d'ajustament per lots (batch adjustment method), ComBat, en una gran mostra de dades del consorci ENIGMA per tal de reduir l'heterogeneïtat relativa al centre i com a conseqüència augmentar-ne el poder estadístic.

4.1.2. MÈTODES

Dades

Per aquest estudi es van utilitzar dades de grossor cortical, àrea superficial i volums subcorticals de 33 centres de l'ENIGMA Schizophrenia Working Group. Concretament, vam usar dades individuals de 2897 pacients diagnosticats amb esquizofrènia (mitjana d'edat de 34 anys, i un 34% de sexe femení) i 3141 controls sans (mitjana d'edat de 33 anys, i un 49% de sexe femení).

Les dades van ser processades amb el programari FreeSurfer[100] seguint els protocols establerts pel consorci ENIGMA, disponibles a https://enigma.ini.usc.edu/protocols/imaging-protocols/ . Per a les regions d'interès corticals, aquests protocols suposaven estimar estadístiques corticals a nivell de vèrtex (cortical vertex-wise statistics), l'extracció dela grossor cortical i l'àrea superficial de 70 regions de l'atles Desikan-Killiany (DK) [101] i els respectius controls de qualitat. Per les regions d'interès subcorticals, es van haver de fer estimacions de volum subcortical i controls de qualitat [102].

Metodologia

Vàrem realitzar comparacions de dades de RM entre individuals amb esquizofrènia i controls sans per tal de trobar la significació estadística, el poder estadístic i el familywise error rate (FWER) dels mètodes RE-Meta, ME-Mega i ComBat-Mega. Concretament, vàrem mirar si el mètode ComBat-Mega augmenta la significació estadística (utilitzant un test de permutacions) i el poder (utilitzant una estratègia de petits subgrups, és a dir repetint vàries vegades les anàlisis amb un nombre petit de centres) en les diferències entre persones amb esquizofrènia i controls sans atenuant els efectes del centre.

Comparació de les dades MRI entre subjectes diagnosticats amb esquizofrènia i controls: Per a RE-Meta primer vam comparar els valors de cada ROI entre pacients i controls mitjançant models lineals estàndard amb edat i sexe com a covariables separadament per cada centre. Llavors vam convertir la diferència trobada a efectes Hedges G i la seva variància. A continuació vam fer un metaanàlisi d'efectes aleatoris per cada efecte Hedges G i vam corregir els valors p amb un mètode de múltiples comparacions de Holm.

Per ME-Mega, vam comparar els valors de les ROI entre grups mitjançant models lineals d'efectes mixtos, amb l'edat i el sexe com a covariables i amb el centre com a factor aleatori. Llavors vam dividir la diferència per la desviació estàndard i vam corregir pel biaix de mostres petites fins a obtenir un efecte Hedges G i la seva variància, i vam corregir també els valors p mitjançant múltiples comparacions de Holm.

Pel ComBat-Mega, primer vam extreure els efectes del centre utilitzant funcions de ComBat (modelant els efectes del diagnòstic, el sexe i l'edat) i comparant els valors de cada ROI entre grups mitjançant un model lineal amb edat i sexe com a covariables. Encara que el mètode ComBat faci servir covariables per estimar millor els efectes del centre, aquest no n'extreu els efectes de les covariables de les dades. Finalment, vam seguir els mateixos passos que pels dos anteriors, vam convertir les diferències a efectes Hedges G i la seva variància, vam corregir els valors p amb el mètode de múltiples comparacions de Holm.

Comparació de la significació estadística: Per fer la comparació estadística, vam fer un test de permutacions basat en el mètode de Draper-Stoneman, ja que és un dels mètodes que millor controla el FWER [103]. Concretament, vàrem assignar aleatòriament els subjectes a pacient o control aleatòriament 1000 vegades, i després vam mirar les diferències entre els p-valors (transformada logit perquè sigui més sensible a p-valors petits).

Avaluació del poder estadístic: Per mirar si el ComBat-Mega incrementava el poder estadístic vam utilitzar una estratègia de petites submostres. Concretament, vam repetir 500 vegades les anàlisis incloent cada vegada només les dades de 10 centres. Llavors vam comptar el nombre de vegades que aquestes anàlisis amb la submostra de 10 centres eren capaces de detectar diferències entre pacients i controls.

Determinació empírica del FWER: utilitzant les dades del test de permutacions anterior, vam comptar la proporció de permutacions en què com a mínim una ROI tenia un p-valor <0,05.

4.1.3. RESULTATS

Comparació de les dades MRI entre subjectes diagnosticats amb esquizofrènia i controls:

De mitjana, amb el ComBat-Mega, els subjectes amb diagnòstic esquizofrènia van mostrar un còrtex més prim i una àrea superficial més petita en quasi totes les ROIs. Aquest grup de pacients també van mostrar de mitjana volums més petits de tàlem bilateral més petit, l'hipocamp, l'amígdala i l'acumbent dret. Volums més grans bilateralment en el ventricle lateral, el putamen i el pàl·lidum.

Comparació de la significació estadística: L'estimació de la mida de l'efecte, calculat en Hedges g, per les diferències va ser similar entre els diferents mètodes, però la significació estadística va ser millor en el ComBat-Mega.

Avaluació del poder estadístic: Fent servir el mètode de petites submostres, el poder estadístic pel ComBat-Mega va ser de 83,5%, superior al RE-Meta (poder estadístic del 53,7% amb Wilcoxon p-valor < 0.001) i al ME-Mega (poder estadístic del 80,4% amb Wilcoxon p-valor <0.001).

Determinació empírica del FWER: El FWER va ser <=0,05 en tots els mètodes (RE-Meta: 0.024; ME-Mega: 0.027; ComBat-Mega: 0.025).

4.1.4. CONCLUSIÓ

En aquest estudi vam testar l'ús del ComBat-Mega en dades del consorci ENIGMA Schizophrenia Working Group. En la comparació entre grups de les dades de MRI els resultats van ser similars a estudis anteriors realitzats amb altres mètodes diferents. L'ús del ComBat va augmentar la significació estadística, amb intervals de confiança més estrets i valors p inferiors, en les diferències entre pacients diagnosticats amb esquizofrènia i controls. Els resultats en el FWER van ser similars en tots els mètodes, i els resultats de l'avaluació de poder estadístic també van ser superiors en el ComBat.

4.2. ARTICLE 1: BIASED ACCURACY IN MULTISITE MACHINE-LEARNING STUDIES DUE TO INCOMPLETE REMOVAL OF THE EFFECTS OF THE SITE

4.2.1. INTRODUCCIÓ

Tal com he mencionat anteriorment, en estudis multicèntrics és comú tenir en compte la procedència del centre per tal d'extreure'n els efectes durant la fase d'entrenament dels models, per exemple amb l'ús del ComBat, que treu l'efecte de les diferències en la mitjana i la variància relatives a les imatges de RM de diferents dispositius o centres. El que no és tan freqüent i que també s'ha de tenir en compte, és que en estimar el rendiment d'un model, encara queden efectes que potencialment poden esbiaixar els resultats, normalment inflant-los, tot i que en algun cas pot reduir-ne falsament l'eficàcia.

En aquest estudi posem alguns exemples per demostrar que és un problema real, i alhora proposem una metodologia per a poder extreure aquests efectes de forma senzilla.

A més, proporcionem un paquet d'R ("multisite.accuracy") per tal de poder extreure aquests efectes de forma fàcil.

4.2.2. MÈTODES

Hem provat dos mètodes diferents de controlar aquests efectes, i que ambdós es troben en el paquet: un mètode basat en metaanàlisis i un altre covariant pel centre.

Un estudi diferent per cada centre (mètode basat en meta-anàlisis)

Aquest primer mètode consisteix a crear un algorisme d'aprenentatge automàtic diferent per cada centre. Aquest enfocament implica tenir suficients dades per cada centre per a poder ajustar correctament un model d'aprenentatge automàtic, cosa difícil en la majoria d'estudis de neuroimatge.

Per a estimar el rendiment els autors només han d'utilitzar les mètriques comunes en estudis d'un sol centre:

- Resposta binària: rendiment balancejat (Balanced Accuracy), sensitivitat i especificitat, l'Àrea sota la curva ROC...
- Resposta continua: correlació entre la predicció del model i el valor real, l'error mitjà al quadrat (MSE) entre variable predita i valor real..
- Estudis de supervivència: Un model de cox per tal d'extreure el perill relatiu (HR), que es pot interpretar com la correlació entre el valor predit de supervivència i el real. HR = 1 vol dir que no hi hauria correlació.

Un cop calculades les mètriques per cada estudi, les combinem utilitzant metaanàlisis per obtenir una sola mètrica conjunta. Per meta-analitzar estimacions de variables binàries entre centres, els analistes primer hauran de convertir les proporcions de rendiment en sensitivitat i l'especificitat, és a dir en variables contínues. Per fer-ho poden usar la transformació logit i arcsin. Nosaltres incloem la transformació logit.

Per a meta-analitzar MSE hem adaptat les fórmules de Nakagawa et al., 2015 [104]. Per a correlacions es pot utilitzar la transformació de Fisher. I per HR es pot transformar amb la funció log.

Estimar el rendiment del model afegint el centre com a covariable (mètode de la covariable)

Aquest mètode, estima una sola mètrica de rendiment per tota la mostra, incloent-hi el centre com a covariable en la fórmula. Els passos detallats per diferents mètriques es poden trobar a l'article original a la propera secció. Per fer simple el resum d'aquest mètode mencionaré de forma breu tres exemples de com n'hem calculat les mètriques:

Sensitivitat:

$$Se = \text{logistic}([\beta_0 \text{ in } m]) \quad \text{on} \quad m = \text{Firth}\left(\left[y_{predita} == 1\right] \sim \mathbf{factor}(\boldsymbol{centre})\right)$$

Error mitjà al quadrat:

$$MSE = \text{mitjana}([e \text{ in } m]^2) \quad \text{on} \quad m = \text{LM}\left(y \sim \text{offset}\left(y_{predita}\right) + \boldsymbol{factor}(\boldsymbol{centre})\right)$$

Correlació:

$$r = \frac{1}{\sqrt{1+\frac{df}{[t_1 \text{ in } m]^2}}} \quad \text{on} \quad m = \text{LM}\left(y \sim y_{predita} + \boldsymbol{factor(centre)}\right)$$

Firth: Regressió logística de biaix reduït de Firth

LM: model lineal

Se: sensitivitat

Y: resposta real

$y_{predita}$: resposta predita

4.2.3. RESULTATS

Dades simulades

Per provar els mètodes hem utilitzat dos grups de dades, el primer consisteix en unes dades simulades. En el cas en què no hi ha efectes reals a les dades, tots els models van retornar un rendiment balancejat al voltant de 0,5 si tampoc no hi havia efectes del centre o era un efecte petit (Figura 8 a l'esquerra). Quan l'efecte del centre era gran, només el mètode de la covariable i del metaanàlisi seguia mostrant un rendiment proper al 0,5, mentre que si no es controlava aquest efecte, s'obtenia un rendiment inflat proper a 0,7.

En el cas en què l'efecte real era molt gran en tots els centres (Figura 8 a la dreta), tots els mètodes van retornar un rendiment al voltant de 0,84 si no hi havia efectes del centre o si eren petits. Quan l'efecte del centre era molt gran, els mètodes seguien amb un rendiment similar, mentre que l'absència de control portava a una reducció del rendiment.

Per tant els mètodes eren capaços de controlar correctament els efectes del centre en les situacions provades.

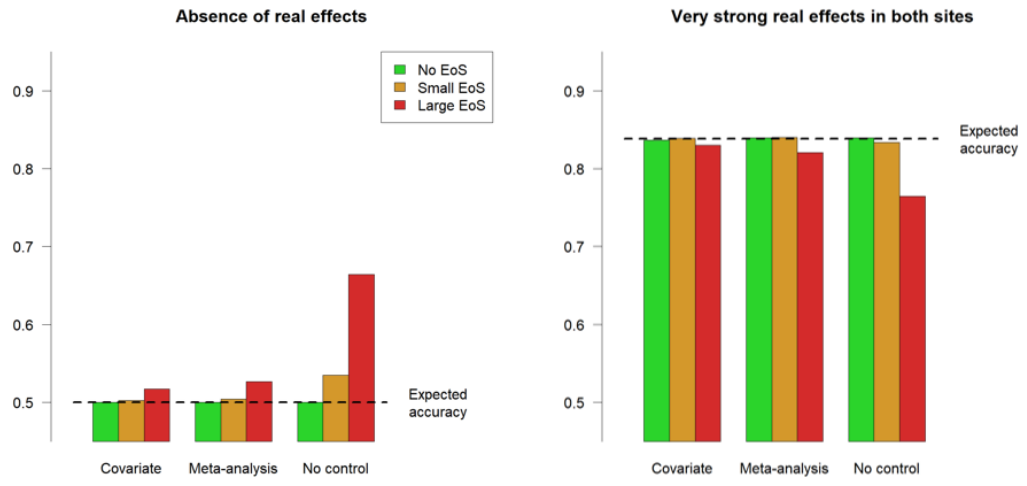


Figura 8. Resultats d'aplicar els diferents mètodes en dades simulades. A l'esquerra un gràfic en el cas que a les dades no hi hagi efectes reals; a la dreta en el cas que hi hagi efectes reals grans en els dos centres.

Dades reals de RM

Per comprovar el comportament dels mètodes en dades reals de RM, vam utilitzar dues bases de dades públiques (OASIS i IXI, https://www.oasis-brains.org/ i https://brain-development.org/)[105]). Després de preprocessar les dades vam definir una resposta que no

fos substancialment diferent entre centres i fortament correlacionada amb les diferències de la RM entre centres. També vam assegurar-nos que no tingués cap relació amb l'edat o el sexe dins de cada centre perquè aquestes variables no eren d'interès aquí. És a dir que en aquestes dades els efectes provenien fortament de les pròpies diferències entre centres.

Seguint un enfocament d'aprenentatge automàtic de validació creuada de 10 folds, vam treure en la fase d'entrenament efectes d'edat, sexe i centre de les dades de RM, i després aplicar aquests estimats a les mostres de test. La resposta va ser calculada com a error mitjà al quadrat.

En aquestes dades, sense controlar l'efecte del centre en estimar el rendiment del model, el model de Machine learning va obtenir un MSE=5.3, estadísticament millor que per atzar (MSE=6.8 i wilcoxon test $p = 0.004$).

Controlant l'efecte del centre, el MSE del model va ser de 1.5, estadísticament pijtor que per atzar (MSE=0.5, wilcoxon test $p = 0<0.001$). Per tant aquest model no era capaç de predir, tal com era esperable, i només afegia soroll.

4.2.4. CONCLUSIÓ

En aquest article, vam mostrar com controlar els efectes del centre en el moment d'estimar el rendiment d'un model. Vam provar dos mètodes diferents, un basat a fer un estudi per separat per cada centre i tractar-ho com un metaanàlisi, i l'altre basat a afegir el centre com a covariable. A més, facilitem un paquet d'R per tal de facilitar-ne el seu ús. També proveïm el codi i totes les instruccions per tal de replicar tots els càlculs.

Aquest estudi demostra que utilitzant qualsevol dels dos mètodes els biaixos del centre eren menors i substancialment més petits que els biaixos observats quan no es controlaven els efectes. Per tant, podem concloure que el control dels efectes del centre en estudis d'aprenentatge automàtic pren importància també en l'estimació del rendiment dels models.

4.3. ARTICLE 2: COMBINING MRI AND CLINICAL DATA TO DETECT HIGH RELAPSE RISK AFTER THE FIRST EPISODE OF PSYCHOSIS

4.3.1. INTRODUCCIÓ

S'han demostrat associacions entre algunes mesures de RM i trastorns mentals, i això obre la porta a la cerca de biomarcadors basats en la RM [106]. Les noves tècniques d'aprenentatge automàtic han incrementat les possibilitats d'obtenir eines que puguin ajudar al clínic, tot i que de moment estiguin lluny de ser perfectes [87]. Aquestes eines podrien ser útils per ajudar en el diagnòstic, la predicció de resposta a un tractament o estimar el risc de tenir una mala evolució, permetent un ajustament de la intervenció segons cada subjecte.

En aquest estudi, hem investigat si les dades estructurals cerebrals de RM podrien ajudar a detectar els pacients amb un primer episodi de psicosi en alt risc de recaiguda (HRR-FEP). Per dur a terme aquesta tasca, hem creat una eina que permet detectar de forma senzilla aquells pacients en HRR-FEP. A més, reportem també els mètodes que hem utilitzat per tal de trobar la configuració òptima de paràmetres d'aprenentatge automàtic relatius a la RM. Per a l'optimització de paràmetres ho hem fet en dues mostres independents exclusives per aquesta tasca.

Finalment, també proporcionem de forma lliure i gratuïta un programari basat en tècniques d'aprenentatge automàtic per tal que altres grups puguin desenvolupar els seus propis mètodes de detecció, i a més proporcionem una pàgina web on poder estimar de forma senzilla el risc HRR-FEP per tal d'ajudar a altres grups a replicar independentment els nostres mètodes.

4.3.2. MÈTODES

Participants

La cohort incloïa 227 pacients amb un primer episodi psicòtic (PEP) de 7 hospitals diferents d'Espanya, alguns provinents d'un estudi multicèntric anterior, seguits prospectivament

durant 2 anys. Vam estimar el grandària mostral fent servir una meta-anàlisi anterior en què la ràtio de recaiguda a dos anys era d'un 37%. Amb aquesta estimació, la mostra necessària per poder detectar un risc de perill (hazard ratio (HR)) de 2 entre pacients amb HRR-FEP i pacients en un risc baix de recaiguda eren 190 subjectes. Per compensar possibles pèrdues de seguiment prematures vam incloure un 20% més de subjectes.

La mitjana d'edat era de 24,2 anys (SD 7.4), 78 eren dones (34,4%).

Vàrem definir una recaiguda com un empitjorament dels símptomes durant com a mínim una setmana i amb almenys un d'entre 8 ítems de l'escala PANSS (P1, P2, P3, N1, N4, N6, G5 I G9) puntuant per sobre de 3 [51]. Per contra, la remissió va ser definida com a puntuar per sota de 3 en tots els 8 ítems abans mencionats. La recaiguda només va ser considerada després de 6 mesos des de la remissió.

Per cada participant vam obtenir una seqüència T1 de RM. Vam utilitzar un pipeline de preprocessat basat en morfometria basada en vòxels (VBM), ja que en un estudi previ havíem detectat millors resultats de rendiment utilitzant dades de VBM [54].

Preprocessat de les dades

Extracció dels efectes del centre: Aquest estudi contenia dades provinents de diferents centres, i això pot incrementar el soroll o donar resultats confusos. Això pot ser per exemple per culpa de les diferències en l'obtenció de les dades de RM entre diferents màquines. Per tal de treure aquests efectes, vàrem utilitzar el mètode ComBat, detallat com a article adjunt de la tesi.

Els paràmetres del mètode ComBat van ser calculats exclusivament utilitzant la mostra d'entrenament, i posteriorment van ser aplicats a la mostra de test.

A més, els efectes del centre d'obtenció també van ser controlats en estimar el rendiment del model (utilitzant el mencionat a "*Article 1: Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site*").

Optimització dels paràmetres dels mètodes d'aprenentatge automàtic basats en RM: Els paràmetres dels mètodes d'aprenentatge automàtics van ser calculats utilitzant dos mostres

de dades independents. La principal raó va ser que volíem evitar qualsevol mena de filtració de dades entre les dades principals de l'estudi i el càlcul dels paràmetres.

La primera mostra contenia 120 subjectes sans, i vam utilitzar la seva RM per predir una variable continua, l'edat. La segona mostra contenia 255 subjectes, la meitat amb diagnòstic d'esquizofrènia i l'altra meitat controls sans, i vam fer servir la seva RM per predir una variable binària, entre si tenien diagnòstic d'esquizofrènia o no.

Els paràmetres per defecte que vam considerar van ser: $\sigma = 4$ mm (que correspon a un FWHM = 9,5 mm) i una mida de vòxel de $3 \times 3 \times 3$ mm³. A partir d'aquests paràmetres vam provar si el rendiment dels models depenia de:

- Afegir més informació:
    - Dades de RM de matèria grisa i blanca
    - Imatges modulades i no modulades, ja que aporten més informació volumètrica
    - El volum global de matèria grisa i el volum global cerebral
    - Anomalies cerebrals centrals (el càvum septum pellucidum i l'absència de adhesion inthertalamica) reportats prèviament com a bons predictors en PEP
- Modificar la mida del kernel de suavitzat (entre $\sigma = 2$ i 6 mm, corresponent a FWHM ≈ 5.3–15.8 mm)
- Utilitzar mètodes d'aprenentatge per conjunt (ensemble learning), que són mètodes que pretenen millorar el rendiment combinant diferents models.
- Augmentar la mida del vòxel, per tal de reduir el cost computacional (3x3x3 mm^3, 6x6x6mm^3 i 12x12x12mm^3)
- Limitar les anàlisis a només vòxels estadísticament significatius

<u>Creació i validació dels models de detecció de HRR-FEP</u>

Vàrem utilitzar un esquema de validació creuada per a crear i validar els models. Això vol dir, que específicament vam dividir la mostra en 10 grups diferents (o folds) intentant preservar un nombre similar de recaigudes a cada fold. Primer vam crear un model utilitzant dades de subjectes dels grups 2 al 10 (la mostra d'entrenament), i vam estimar el risc de recaiguda sobre els subjectes del grup 1 (la mostra de test). Successivament vam repetir el procediment

posant com a mostra de test cada un dels diferents grups. D'aquesta manera mai vàrem utilitzar els mateixos subjectes per a crear el model que per validar-lo.

Per crear els models, en els passos d'entrenament vam ajustar un model de regressió múltiple.

Variable dependent: temps fins a la recaiguda.

Variables independents: Dades clíniques (escales PANSS, GAF, MADRS, YMRS, el diagnòstic i si el pacient prenia un tractament d'antipsicòtics injectables de llarga durada o no) i els valors de les dades de RM pre-processades.

Abans de la regressió vam treure els efectes d'edat i sexe de les dades de RM utilitzant models lineals estàndard. Igual que tots els passos descrits anteriorment, l'efecte del sexe i l'edat va ser estimat en la mostra d'entrenament i aplicat a la mostra de test de forma separada. Vam escalar les variables clíniques a un rang de [0-1] per tenir una distribució similar a la dels vòxels. Per evitar el sobreajustament vam utilitzar la regressió de Lasso (descrit a la introducció 1.5.1 Regressió de Lasso), que automàticament selecciona uns quants regressors penalitzant la suma del valor absolut dels coeficients i que ha estat demostrat que és capaç de manejar dades amb una alta dimensionalitat i tot i així aconseguir models amb un alt rendiment [107]. El valor de regularització defineix el grau de penalització, entre 0 (cap penalització, com a una regressió lineal simple) i infinit (màxima penalització). Aquest paràmetre va ser trobat mitjançant una validació creuada interna dins de la mostra d'entrenament. Tots aquests paràmetres estimats en la mostra d'entrenament, posteriorment han estat aplicats a la mostra de test per tal de validar el rendiment del model.

Resumint, vàrem trobar un model d'estimació de risc en la mostra d'entrenament, i posteriorment ho vàrem aplicar en una mostra de test per trobar el seu risc de recaiguda.

Per a determinar si un pacient estava considerat com a pacient de risc HRR-FEP vam multiplicar cada coeficient del model de lasso pel valor de la variable del pacient en qüestió i finalment vam sumar aquestes multiplicacions. Si la suma era >0 (corresponent a un HR > 1) vam considerar que el pacient estava en alt risc de recaiguda (HRR-FEP). Per contra, si la suma era igual o inferior a 0 (corresponent a un HR<=1) vam considerar el pacient en baix risc de recaiguda.

Finalment, per mirar si els pacients en risc de HRR-FEP tenien estadísticament més recaigudes que els pacients en baix risc, vam utilitzar el paquet d'R "multisite.accuracy" creat i publicat arrel del primer article de la tesi (4.2 Article 1: Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site). Aquest paquet té en compte l'efecte dels diferents centres per a estimar el rendiment del model. Concretament vam utilitzar un model de regressió de perills proporcionals de Cox d'efectes mixtos (mixed-effects Cox proportional hazards regression). La variable dependent va ser el temps fins a la recaiguda i la variable independent va ser el risc estimat (alt risc de recaiguda vs. Baix risc de recaiguda) i el centre va ser considerat un efecte aleatori de no interès.

Eines disponibles

Hem facilitat una web on altres centres poden afegir les dades dels seus pacients per tal de calcular amb el nostre model el risc de recaiguda (https://www.mripredict.com/hrr-fep/).

També hem creat un software (un paquet d'R i una interfície gràfica per facilitar-ne l'ús) que pot ser utilitzat per altres investigadors per a crear els seus models amb els seus subjectes i les seves pròpies preguntes d'investigació, disponible gratuïtament a https://mripredict.com.

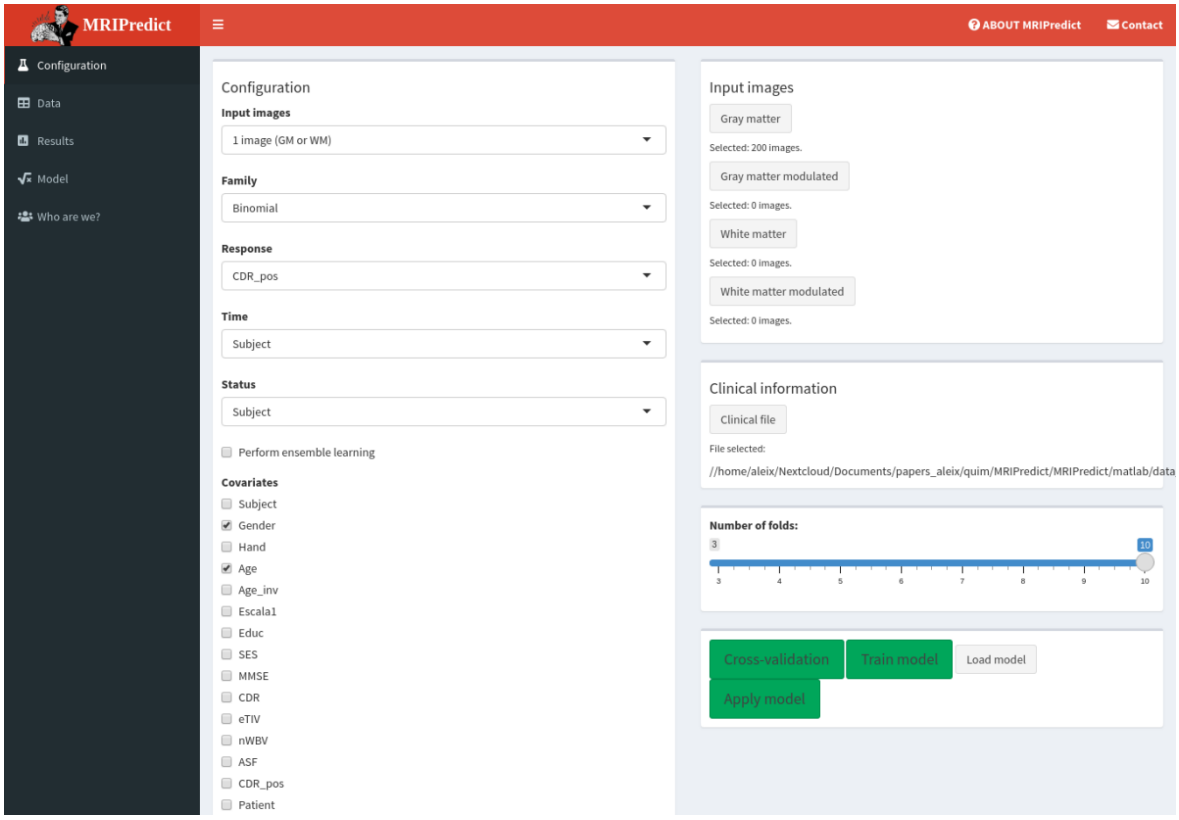


Figura 9. Software MRIPredict. Pàgina on definir el model, les variables i les seves característiques.

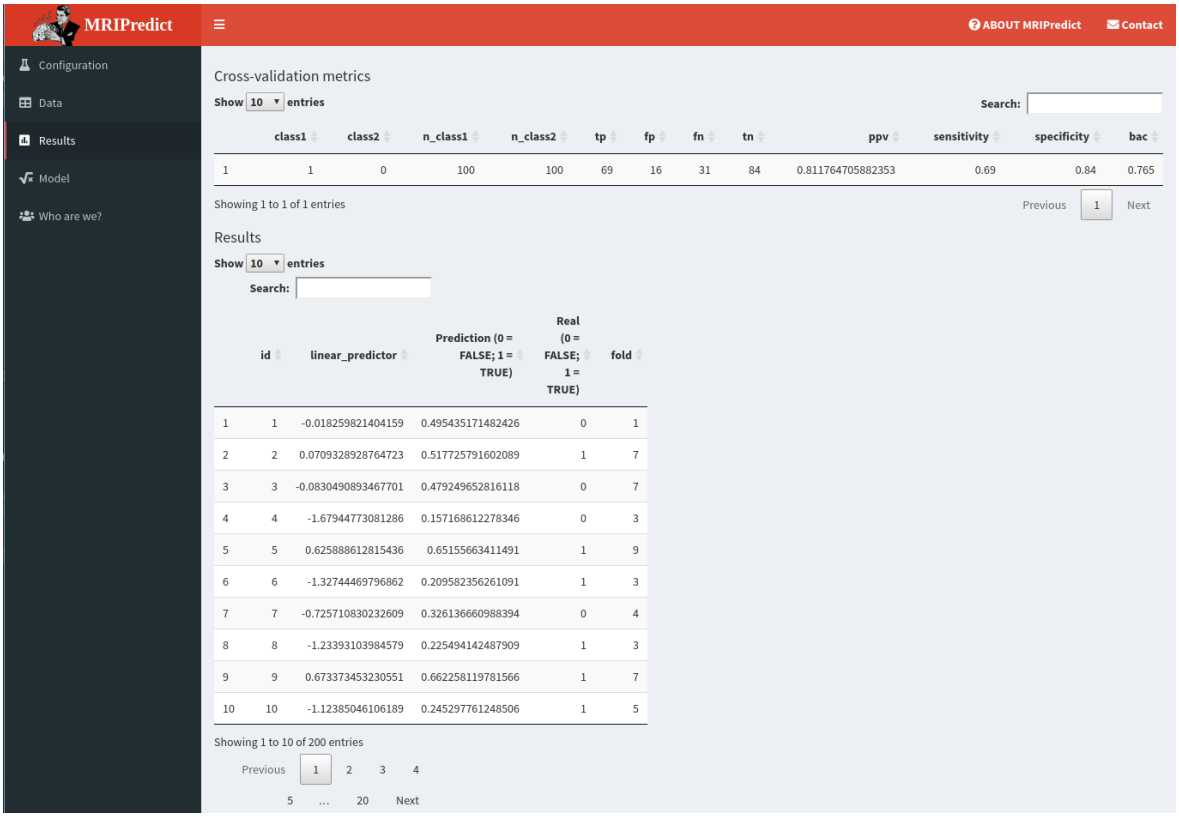


Figura 10. Software MRIPredict. Pàgina on veure els resultats d'una validació creuada del model.

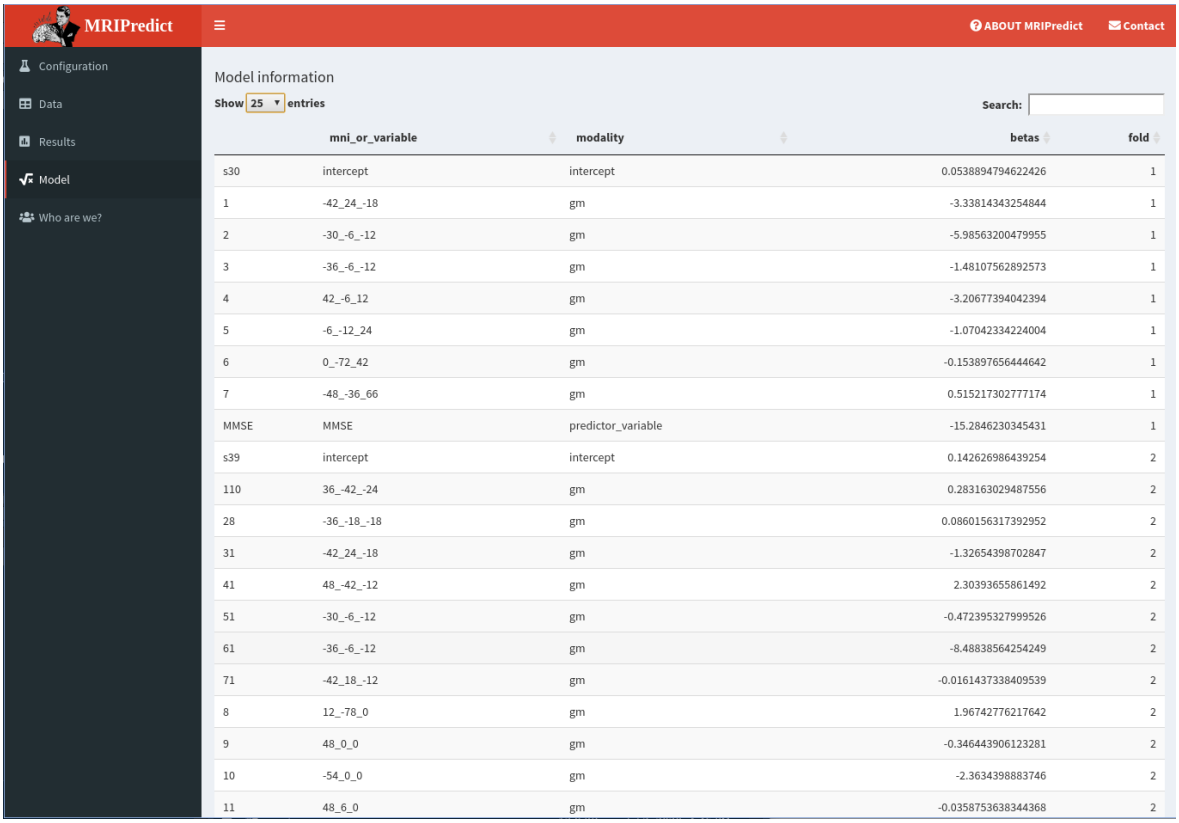


Figura 11. Software MRIPredict. Pàgina on veure els models utilitzats junt amb quines variables s'han utilitzat.

4.3.3. RESULTATS

Fase d'optimització de paràmetres

En la fase d'optimització vam trobar que la millor combinació de dades i paràmetres per a la predicció era l'ús de dades de RM de substància grisa, blanca, modulada grisa, modulada blanca i l'ús de tècniques d'ensemble learning. Els altres paràmetres provats no van millorar respecte dels per defecte. Sí que vam optar per incrementar la mida dels vòxels a 12x12x12mm^3 per reduir el cost computacional i perquè el rendiment en les mostres d'optimització van donar resultats similars a la mida original de vòxel.

Detecció de pacients en risc de recaiguda

La regressió de Cox del temps de recaiguda comparant pacients classificats com a HRR-FEP vs. baix risc de recaiguda va donar resultats clínicament rellevants (HR = 4,58), és a dir que els pacients classificats com a HRR-FEP tenien vora 5 vegades més de risc de recaiguda que els classificats amb risc baix, amb una significació estadística límit amb una p = 0.048, Z=1,98, HR 95% interval de confiança = 1,01-20,74.

En els 114 pacients classificats com a subjectes amb alt risc de recaiguda hi va haver 13 recaigudes reals, representant una ràtio de recaigudes del 14,8% al cap de 24 mesos. En canvi, entre els 113 subjectes classificats amb risc baix de recaiguda només hi va haver 3 recaigudes, que representa una ràtio de recaigudes del 2,9% al cap de 24 mesos.

Vam calcular que el poder estadístic d'obtenir un HR de 4,58 amb 16 recaigudes a la mostra era de 72%.

Les variables seleccionades automàticament per la regressió de Lasso per crear l'eina de detecció HRR-FEP van ser: el diagnòstic de trastorn esquizoafectiu, la dificultat en el pensament abstracte i un mal control dels impulsos, i l'augment o disminució de substància grisa i blanca no-modulada i modulada en diverses regions cerebrals (el model exacte es pot trobar a l'article original adjuntat a la següent secció).

4.3.4. CONCLUSIÓ

En aquest article vam crear una eina per tal de detectar els pacients amb un primer episodi psicòtic en alt risc de recaiguda utilitzant una cohort de 227 subjectes. El model va mostrar índex de rendiment satisfactori per detectar els subjectes HRR-FEP. El risc de recaiguda va ser 4,58 vegades superior en els individus classificats com a HRR-FEP.

Hem facilitat l'ús del model creat en aquest estudi mitjançant una plataforma web, i hem obert gratuïtament el nostre codi en forma de paquet i interfície de fàcil ús sota la web https://mripredict.com.

Aquesta eina, que en estudis posteriors hauria de ser validada i ampliada sobre una mostra més gran de pacients, podria ser una eina útil en un futur per al clínic, ja que per establir els pacients que es troben en HRR-FEP els metges solen necessitar vàries recaigudes. Aquesta eina podria, per tant, ajudar a ajustar el tractament del pacient, sempre seguint el principi "el primer és no perjudicar", car en la nostra mostra el 85% dels pacients estimats com a HRR-FEP no van recaure. No obstant això, aquest estudi pot servir per a demostrar la potencial utilitat clínica d'una eina basada en aprenentatge automàtic sobre dades de RM per tal d'ajudar al clínic proveint-lo d'una informació addicional.

5. ARTICLES INCLOSOS EN LA TESI

Aquesta tesi està composta per dos articles, i un article adjunt[2] els quals estan inclosos a continuació:

- Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA (Article adjunt) [91]
    - DOI: 10.1016/j.neuroimage.2020.116956
    - Publicat a NeuroImage; Factor d'impacte l'any 2020: 7

- Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site [92]
    - DOI: 10.1016/j.pscychresns.2021.111313
    - Publicat a Psychiatry Research: Neuroimaging; Factor d'impacte l'any 2021: 3.2

- Combining MRI and clinical data to detect High Relapse Risk after the First Episode of Psychosis [96]
    - DOI: 10.1038/s41537-022-00309-w
    - Publicat a NPJ Schizophrenia; Factor d'impacte l'any 2021: 5.04

[2] L'article adjunt (Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA[91]) no forma part dels articles estructurals de la tesi per qüestions burocràtiques. Però es troba com a material adjunt per la seva estreta relació amb el contingut i l'importància que té per a l'article principal (Combining MRI and clinical data to detect High Relapse Risk after the First Episode of Psychosis[96]).

5.1. ARTICLE 1: BIASED ACCURACY IN MULTISITE MACHINE-LEARNING STUDIES DUE TO INCOMPLETE REMOVAL OF THE EFFECTS OF THE SITE

Psychiatry Research: Neuroimaging 314 (2021) 111313

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

journal homepage: www.elsevier.com/locate/psychresns

Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site

Aleix Solanes [a,b], Pol Palau [c,d], Lydia Fortea [a,e,f], Raymond Salvador [c,e], Laura González-Navarro [g], Cristian Daniel Llach [a,e,f,h], Marc Valentí [a,e,f,h], Eduard Vieta [a,e,f,h], Joaquim Radua [a,e,i,j,*]

[a] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
[b] Department of Psychiatry and Forensic Medicine, Autonomous University of Barcelona, Barcelona, Spain
[c] FIDMAG Research Foundation, Barcelona, Spain
[d] CASM Benito Menni Granollers-Hospital General de Granollers, Barcelona, Spain
[e] Biomedical Network Research Centre on Mental Health (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain
[f] Institute of Neurosciences, University of Barcelona, Barcelona, Spain
[g] Faculty of Biology, University of Barcelona, Barcelona, Spain
[h] Barcelona Bipolar Disorders and Depressive Unit, Institute of Neurosciences, Hospital Clinic, Barcelona, Spain
[i] Department of Psychosis Studies, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, United Kingdom
[j] Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

ARTICLE INFO

Keywords:
Bias
Effects of the site
Machine learning
Magnetic resonance imaging

ABSTRACT

Brain MRI researchers conducting multisite studies, such as within the ENIGMA Consortium, are very aware of the importance of controlling the effects of the site (EoS) in the statistical analysis. Conversely, authors of the novel machine-learning MRI studies may remove the EoS when training the machine-learning models but not control them when estimating the models' accuracy, potentially leading to severely biased estimates. We show examples from a toy simulation study and real MRI data in which we remove the EoS from both the "training set" and the "test set" during the training and application of the model. However, the accuracy is still inflated (or occasionally shrunk) unless we further control the EoS during the estimation of the accuracy. We also provide several methods for controlling the EoS during the estimation of the accuracy, and a simple R package ("multisite.accuracy") that smoothly does this task for several accuracy estimates (e.g., sensitivity/specificity, area under the curve, correlation, hazard ratio, etc.).

1. Introduction

When an individual undergoes MRI brain scanning in different MRI devices, the resulting images differ (Focke et al., 2011). These effects of the site (EoS) are relevant because, in many analyses, they may play a critical confounding role. To put a simple example, we might consider a two-site study investigating the effects of a disorder's severity on gray matter volume. Suppose that the sites are imbalanced: one site has individuals with mainly mild forms of the disease, and the other site has individuals with severe forms of the illness. In that case, the observed differences in gray matter volume between individuals with severe and mild conditions in the overall study could indeed represent differences between the two MRI scanning devices. Fortunately, researchers are very aware of the importance of the site's potentially confounding effects and the necessity of controlling them when conducting statistical analyses with tools such as ComBat, which removes differences in mean and variance related to the use of different MRI devices or sites (Radua et al., 2020).

However, analysts do not always control the EoS *when estimating the accuracy* in novel machine-learning multisite MRI studies (Beheshti et al., 2017; Gill et al., 2020; Leger et al., 2020). For example, Archer et al. (2019) removed the EoS from both the "training set" and the "test set" when training the machine-learning model and using it to predict the outcomes with the ComBat tool (Fortin et al., 2017; Radua et al., 2020). Still, they estimated the machine-learning model's accuracy as the area under the curve without controlling the EoS.

* Corresponding author. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), C/. Rosselló, 149, 08036 Barcelona, Spain.
E-mail address: radua@clinic.cat (J. Radua).

https://doi.org/10.1016/j.pscychresns.2021.111313
Received 4 November 2020; Received in revised form 24 April 2021; Accepted 27 May 2021
Available online 29 May 2021

Some readers might argue that such control is unnecessary in studies aiming to classify patients vs. controls with balanced sample sizes. Effectively, if the number of patients and controls is the same within each site, the EoS cannot confound the accuracy's estimation. However, we would like to add that this perfect balance of the outcome across sites is uncommon in multisite studies. The different sites usually include individuals of different ages, varying severity of the disorder, and etcetera.

Other readers might argue that there is no need to control the EoS *when estimating the machine-learning models' accuracy* if the modelers have already removed these effects when training the machine-learning models and using them to predict outcomes. This argument would be correct if we could entirely remove all the EoS. Or at least any EoS potentially usable by the subsequent machine-learning algorithm. Unfortunately, we should not make this assumption. Imagine, for example, that the interaction between two adjacent voxels is different in different sites due to the use of different scanner resolutions or the presence of different spatially correlated noise. The removal of these EoS is not straightforward, and the subsequent machine learning algorithm could use them to "fraudulently" increase the accuracy of the predictions. Note that many machine-learning algorithms are very sophisticated and may potentially detect complex EoS that we may have failed to remove before applying the machine learning algorithm. Glocker et al. (2019) described that even after careful pre-processing with state-of-the-art neuroimaging pipelines, machine learning algorithms easily detected the site. At this point, some readers may wonder whether, as far as machine learning models are so good at detecting the EoS, it would be sensible to use a machine learning model to remove them. However, detection and removal are not the same. The removal of complex EoS may be substantially more complicated than its detection and may distort the data preventing the detection of real effects. For example, the detection of different correlations between adjacent voxels across sites is straightforward using a linear model, while its removal is more complicated and may distort the values of the voxels.

This brief report shows how to control these effects in different circumstances and provide an R package to do this task smoothly ("multisite.accuracy"). With this package freely available, we find no reasons to rely on the strong assumption that a pre-processing pipeline can entirely remove the EoS. We also conduct simulations with a toy multisite machine-learning study and describe an example with real MRI data. In these examples, we remove the EoS from both the "training set" and the "test set." Still, the model's accuracy is biased unless we control the EoS during the estimation of the accuracy.

2. How to control the EoS when estimating the accuracy

This section will encompass two main methods to control the EoS for several accuracy statistics, each with strengths and limitations. We have incorporated both approaches in the new "multisite.accuracy" R package and recommend the use of both to check that they return similar accuracy estimates.

2.1. Separate study for each site ("meta-analysis method")

One way to control the EoS is to create a separate machine-learning model for each site. A strength of such an approach is that the manuscript may report each site's machine-learning model and its accuracy, which may shed light on replicability. However, this method requires that each site's sample size must be large enough to appropriately fit a machine-learning model, a luxury out of reach of most neuroimaging studies.

To estimate a site's machine-learning model's accuracy, the analysts can use simple standard formulas for single-study studies (Table 1, first column).

When the outcome is binary (e.g., responder vs. non-responder), the analysts may calculate the sensitivity as the proportion of responders predicted to respond, the specificity as the proportion of non-responders predicted not to respond, and the balanced accuracy (BAC) as the average of sensitivity and specificity. Note that the BAC is more accurate than simple accuracy (proportion of the overall sample correctly predicted to respond or not respond) if the numbers of responders and non-responders are different in any site. They may also calculate the area under the ROC curve (AUC), which summarizes sensitivity and specificity at different thresholds.

For continuous outcomes (e.g., the severity of the symptoms at follow-up), the analysts may calculate the correlation between the predicted outcome and the actual outcome, or the mean squared error (MSE) between the outcomes (i.e., the average of the squared differences between predicted and actual outcomes). For survival, time-to-event data, they may conduct a simple Cox model (CM) to derive the hazard ratio (HR), which one may interpret as the correlation between the predicted survival and the actual survival (although exponentiated: HR

Table 1
Formulas for estimating accuracy.

	Single-site studies Simple formulas	Model-based formulas	Multi-site studies
Binary outcome			
Sensitivity (for $y = 1$)	$Se = \frac{\text{sum}([y_{predicted} == 1])}{n}$	$Se = \text{logistic}([\beta_0 \text{ in } m])$ where: $m = \text{GLM}_{logistic}([y_{predicted} == 1] \sim 1)$	$Se = \text{logistic}([\beta_0 \text{ in } m])$ where: $m = \text{Firth}([y_{predicted} == 1] \sim \text{factor}(site))$
Specificity (for $y = 0$)	$Sp = \frac{\text{sum}([y_{predicted} == 0])}{n}$	$Sp = \text{logistic}([\beta_0 \text{ in } m])$ where: $m = \text{GLM}_{logistic}([y_{predicted} == 0] \sim 1)$	$Sp = \text{logistic}([\beta_0 \text{ in } m])$ where: $m = \text{Firth}([y_{predicted} == 0] \sim \text{factor}(site))$
Area under the curve	(no simple formula)	Area under the ROC curve	Area under the ROC curve adjusted by site covariate
Continuous outcome			
Mean squared error	$MSE = mean(e^2)$ where: $e = y - y_{predicted}$	$MSE = mean([e \text{ in } m]^2)$ where: $m = \text{LM}(y \sim \text{offset}(y_{predicted}))$	$MSE = \text{mean}([e \text{ in } m]^2)$ where: $m = \text{LM/LMM}(y \sim \text{offset}(y_{predicted}) + \text{factor}(site))$
Correlation	$r = \text{cor}(y_{predicted}, y)$	$r = \frac{1}{\sqrt{1 + \frac{df}{[t_1 \text{ in } m]^2}}}$ where: $m = \text{LM}(y \sim y_{predicted})$	$r = \frac{1}{\sqrt{1 + \frac{df}{[t_1 \text{ in } m]^2}}}$ where: $m = \text{LM/LMM}(y \sim y_{predicted} + \text{factor}(site))$
Survival outcome			
Cox regression	(no simple formula)	$HR = \exp([\beta_1 \text{ in } m])$ where: $m = \text{CM}(y \sim y_{predicted})$	$HR = \exp([\beta_1 \text{ in } m])$ where: $m = \text{CM/CMM}(y \sim y_{predicted} + \text{factor}(site))$

AUC: area under the curve; BAC: balanced accuracy; CM: Cox model; CMM: Cox mixed model; Firth: Firth's bias-reduced logistic regression with factor "site" coded to sum zero; GLM: generalized linear model; HR: hazard ratio; LM: linear model; LMM: linear mixed model; MSE: mean squared error; ROC: receiver operating characteristic; Se: sensitivity; Se: specificity.

= 1 means no correlation).

Afterward, he/she might combine these accuracy estimates in a single overall accuracy using meta-analysis (Table 2). To meta-analyze binary accuracy estimates of the different sites, the analysts must first convert the proportions (sensitivity and specificity) into continuous variables, using methods such as the logit, arcsine, and the Freeman-Tukey double arcsine transformations. Schwarzer and colleagues do not recommend the latter because the back-transformation of the meta-analytic estimate into a proportion can lead to inconsistent results (Schwarzer et al., 2019). We include the logit transformation because it is also used in Firth's bias-reduced logistic regression described next. Still, the arcsine transformation would be a good alternative. For meta-analyses of MSE, we adapted the formulas provided by Nakagawa et al. (2015). A standard method to meta-analyze correlations is the Fisher transformation. Finally, one can combine HR by transforming them with the log function.

2.2. Estimating accuracy covarying for the site ("covariate method")

The inclusion of a site-covariate makes estimating accuracy more difficult but deriving the appropriate formulas may still be relatively straightforward. We, for example, used the following trick (Table 1). First, we converted the simple formulas to model-based formulas. We used intercept-only models of the success or error for sensitivity and specificity. In contrast, for MSE, correlation, and CM, we used models in which the independent variable was the predicted outcome and the dependent variable the actual outcome. For the correlation, we also used the inverse of the standard function to derive the t statistic from r. Second, we converted these models into linear models adding the site as a factor.

Therefore, we propose using an intercept-only logistic regression for sensitivity and specificity. The dependent variable is a successful prediction of the event, and the site is a factor coded to sum zero. In the simulations described in Section 3.1, we found that standard logistic regression often yielded implausible accuracy estimates when the machine-learning model had predicted that all patients of a site would respond to the treatment or that none of them would. For this reason, we suggest replacing standard logistic regression with Firth's bias-reduced logistic regression (Firth, 1993), which Heinze and Schemper (2002) proposed as an ideal solution to the problem of separation in logistic regression.

Table 2
Transformations to meta-analyze accuracy estimates from different sites.

	Transformation	Variance of the transformed outcome	Back-transformation of the meta-analytic estimate
Binary outcome			
Specificity (for $y = 0$)	$logitSp = \text{logit}(Sp)$	$\text{var}(logitSp) = \frac{1}{Sp \cdot (1 - Sp) \cdot n}$	$Sp = \text{logistic}(logitSp)$
Sensitivity (for $y = 1$)	$logitSe = \text{logit}(Se)$	$\text{var}(logitSe) = \frac{1}{Se \cdot (1 - Se) \cdot n}$	$Se = \text{logistic}(logitSe)$
Area the under curve	(not needed)	(variance of the AUC)	(not needed)
Continuous outcome			
Mean squared error	$logRMSE = \log(\sqrt{MSE}) + \frac{1}{2 \cdot (n - 1)}$	$\frac{1}{2 \cdot (n - 1)}$	$MSE = \exp(2 \cdot logSE)$
Correlation	$zFisher = \text{atanh}(r)$	$\text{var}(zFisher) = \frac{1}{n - 3}$	$r = \tanh(zFisher)$
Survival outcome			
Cox regression	$logHR = \log(HR)$	(variance returned by the cox regression)	$HR = \exp(logHR)$

AUC: area under the curve; HR: hazard ratio; MSE: mean squared error; RMSE: root mean square error; Se: sensitivity; Se: specificity.

Several methods estimate the area under a ROC curve (ROC) adjusted for covariates, such as the nonparametric kernel-based method (Rodriguez-Alvarez et al., 2011) or the semiparametric approach(Janes and Pepe, 2009). For the R package, we have used the latter because it accepts the use of categorical covariates.

To estimate the error of predicting a continuous variable, we propose using the residuals of a linear model (LM) or a linear mixed model (LMM). The dependent variable is the actual outcome, the predicted outcome is an offset, and the site is a fixed-effects factor for LM and a random-effects factor for LMM. For correlations, we propose using an LM in which the predicted outcome is the independent variable. Finally, we suggest using a CM or Cox mixed model (CMM) for survival outcomes in which the dependent variable is the survival curve. The independent variable is the log predicted HR, and the site is a fixed-effects factor for CM and a random-effects factor for CMM.

We cannot offer a definitive guide about choosing mixed- or fixed-effects models. Mixed-effects models should estimate the intercept of small sites better, but they sometimes fail to converge, and the difference between them is often negligible. In the R package, the function uses fixed-effects models by default, but the user can ask a mixed-effects model if he/she wishes. In the latter case, we have coded that the function still uses a fixed-effects model if the mixed-effects model fails or returns any warning. Besides, the user may ask the function to use the "meta-analysis" method and check that the methods yield similar accuracy estimates.

2.3. Examples of these methods

In the Supplement, we include examples of accuracy estimation with the methods presented here in the presence and absence of EoS and real effects. The examples show that the methods efficiently avoid the EoS-related bias in accuracy while not modifying the accuracy due to real effects.

In the absence of real effects and EoS, all methods behaved similarly: they return a BAC and AUC around 0.5 (i.e., as tossing a coin), a ratio of MSE using the predictions to the MSE using the mean around 2 (i.e., predicting with the mean is twice more accurate than using the machine learning predictions), a correlation around 0, and HR around 1 (i.e., no relationship between the predictions and the outcome).

In the presence of real effects but not EoS, all methods behaved again similarly: BAC around 0.75, AUC around 0.85, MSE ratio around 0.6, correlation around 0.7, and HR about 2.5.

In the presence of EoS but not real effects, the covariate method (Section 2.2) and the meta-analysis method (Section 2.1) behaved like if there were no EoS. Simultaneously, the lack of control of the EoS led to inflated accuracy: BAC around 0.75, AUC around 0.8, correlation around 0.5, and HR about 1.7.

Finally, in the presence of both real effects and EoS, the covariate method and the meta-analysis method mostly behaved as if there were no EoS. At the same time, the lack of control of the EoS led again to inflated accuracy.

3. A toy multisite machine-learning study

3.1. Simulations

We may imagine a multisite study in which we aim to use a patient structural brain MRI to predict whether he/she will respond to a treatment or not. For simplicity, we will imagine that we have only two

equal-sized sites, that the brain MRI has only two voxels, and that the machine-learning model is just a logistic regression.

We will simulate that the values of two voxels and their interaction will always differ between sites, which is realistic given the devices' known effects on the MRI signals and the massive number of voxels in the brain. However, these differences in voxels' values do not involve EoS unless there are also imbalances in the outcome across sites. We will simulate different levels of the following parameters:

a) The presence of EoS, creating imbalances in the outcome between the two sites. We will simulate three scenarios. First, we will simulate that the percentage of patients responding to the treatment is similar across sites, specifically 50% in site 1 and 50% in site 2, thus preventing any EoS. We will call this scenario "no EoS." Second, we will simulate that the percentage of patients responding to the treatment is slightly different across sites, specifically, 60% in site 1 and 40% site 2, thus potentially creating slight EoS. We will call this scenario "small EoS." Finally, we will simulate that the percentage of patients responding to the treatment is substantially different across sites, specifically, 80% in site 1 and 20% in site 2, potentially creating severe EoS. We will call this scenario "large EoS."
b) The presence of real effects, adding the value of the outcome to the voxels. We will simulate five scenarios. First, we will simulate that there are no real effects, i.e., in neither site 1 nor site 2, there is any relationship between the voxels and the response to the treatment. Second, we will simulate real effects, either moderate or very strong, and present in either both sites or only one site. We will thus have the following scenarios: "absence of real effects," "moderate real effects in both sites," "very strong real effects in both sites," "moderate real effects in only one site," and "very strong real effects in only one site."

We will simulate that the researchers train the machine-learning model using conventional steps: a) they find the EoS in the "training set" (linear regressions in which the dependent variables are the voxel values and the independent variable is the site); b) they remove the EoS from the "training set" (subtracting the EoS predicted by the regression from the voxel values); and c) they adjust the machine learning model in the "training set" (logistic regression in which the dependent variable is the outcome and the independent variables are the voxels values and their interaction). We acknowledge that other methods may be more efficient in removing the EoS in real data. However, we only wanted to simulate the realistic scenario in which the removal of EoS from voxel values is incomplete.

We will then simulate that the researchers predict the response in the patients of an independent "test set" using conventional steps: a) they remove the EoS from the "test set" (subtracting the EoS predicted by the EoS regression fitted above from the voxel values); and b) they apply the machine learning model in the "test set" to predict the response. The reason to regress out the site in both the "training set" and the "test set" is to show that removing EoS from both sets is insufficient to remove possible inflation of model accuracy due to EoS.

Finally, we will estimate the BAC of the predictions using the three site methods of the R package: "covar" (i.e., using covariates as described in Section 2.2 and Table 1), "meta" (i.e., using meta-analysis as described in Section 2.1 and Table 2), and "none" (i.e., not controlling for the EoS).

We conducted these simulations 5000 times. The reader can find the code of the simulations in the Supplement.

3.2. Results

Figs. 1-3 show the median accuracy estimates across the 5000 simulations for each combination of settings.

In the absence of real effects, all methods returned a BAC of around 0.5 if there were no EoS or they were small (**Fig. 1**). If there were large EoS, only the covariate and the meta-analysis methods still yielded a BAC of about 0.5; the lack of control of the EoS yielded a severely inflated BAC.

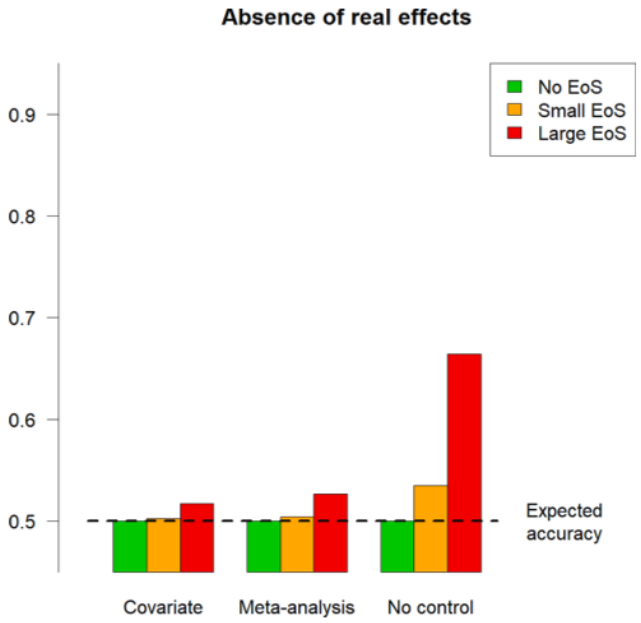


Fig. 1. Estimating the accuracy of the toy multisite machine-learning study using the covariate and meta-analysis methods for controlling for the effects of the site (EoS) or not controlling them in the absence of real effects.

In the presence of moderate real effects in both sites, all methods returned similar BACs around 0.71 independently of the existence of EoS. In other words, the EoS did not significantly bias the estimation of the accuracy (**Fig. 2 bottom**). In the presence of very strong real effects in both sites, all methods returned a BAC around 0.84 if there were no EoS or they were small (**Fig. 2 bottom**). If the EoS were large, the covariate and the meta-analysis method returned a similar BAC, while the lack of control of the EoS yielded a substantially shrunk BAC.

Finally, in the presence of real effects in only one site, all methods returned similar BACs around 0.58 (for moderate real effects) or 0.67 (for very strong real effects) if there were no EoS or they were small (**Fig. 3**). In the presence of large EoS, the covariate and the meta-analysis methods still returned a BAC around 0.58 or 0.67. In contrast, the lack of control of the EoS yielded severely inflated BACs.

4. An example with real MRI data

In this section, we exemplify the same issue using real MRI data. We conducted a machine-learning study using data from two different sites that have apparent differences.

4.1. MRI data

We retrieved the MRI data from two different internet-available datasets (OASIS and IXI, https://www.oasis-brains.org/ and http s://brain-development.org/, (Marcus et al., 2007)). To allow other researchers to reproduce our results quickly, we only included 30 participants from each dataset (see Supplement). Before the statistical analyses, we used SPM12 to align the images with the MNI template, segment them into gray matter, white matter, and cerebrospinal fluid, and spatially normalize the segments to the MNI space. We subsampled the images (voxel size: 32 × 32 × 32 mm) and discarded those voxels with an average gray matter volume of 0.001 or less to reduce the dimensionality.

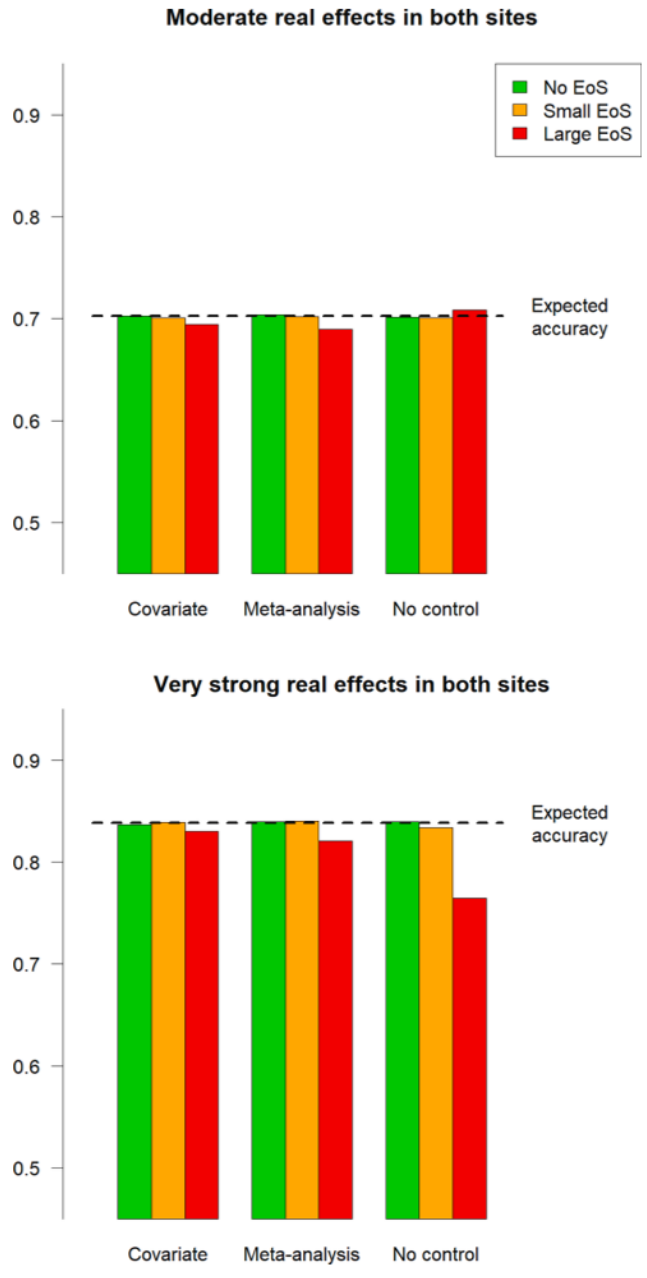


Fig. 2. Estimating the accuracy of the toy multisite machine-learning study using the covariate and meta-analysis methods for controlling for the effects of the site (EoS) or not controlling them in the presence of real effects in both sites.

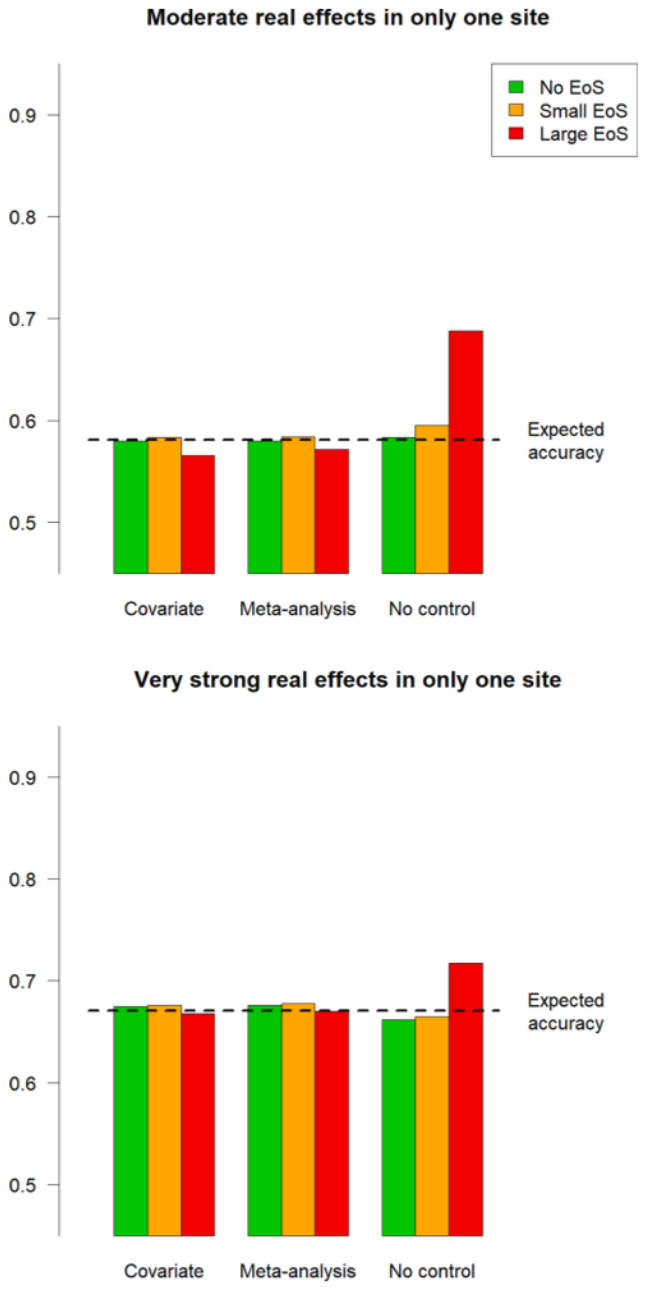


Fig. 3. Estimating the accuracy of the toy multisite machine-learning study using the covariate and meta-analysis methods for controlling for the effects of the site (EoS) or not controlling them in the presence of real effects in only one site.

4.2. Simulated outcome

To better exemplify the problem, we simulated a continuous outcome (e.g., decrease in disorder severity after treatment) that was substantially different between the sites and strongly correlated with the between-site MRI differences. We forced the outcome to have a null relationship with age or sex within each site because these variables are of no interest here. Thus, we attempted to minimize any potential effect of them. We detail the simulated outcome in the Supplement.

4.3. Machine-learning study

We conducted ten-fold cross-validation. Specifically, we iteratively divided the overall sample into a large "training set" consisting of 54 subjects and a small "test set" of 6 other individuals. Creation of the model in the "training sets" included: a) separately for each voxel, finding the effects of sex, age, and site by fitting a standard multiple regression in which the dependent variable was the value of the voxel, and the independent variables were sex, age, and site; b) separately for each voxel, removing the effects of sex, age, and site by subtracting the predicted value of the voxel according to the standard multiple regression; and c) estimating the machine-learning model by fitting a lasso regression in which the dependent variable was the outcome, and the independent variables were the values of the voxels and their interactions. Note that we removed the effects of sex and age from voxel values to mimic standard practice, but these variables are of no interest

here. We used the package "glmnet" for R (Friedman et al., 2010; R Core Team, 2020) to fit the machine-learning model. Application of the model in the "test sets" included: a) separately for each voxel, removing the effects of sex, age, and site by subtracting the predicted value of the voxel; and b) predicting the outcome with the lasso regression.

We calculated the MSE of the predictions as the mean of the squared differences between the predicted and the actual outcomes. To know whether the predictions were better than chance, we also calculated the MSE of a null model consisting of predicting the outcomes in the "test set" as merely the mean of outcomes in the "training set." To statistically compare whether the machine-learning models' squared errors were smaller than the null model's squared errors, we used a paired Wilcoxon test. Finally, we also calculated the MSE of the predictions controlling the EoS.

4.5. Results

Without controlling the EoS when estimating the accuracy, the machine-learning model's MSE was 5.3, statistically significantly better than the MSE of the null model (MSE=6.8, Wilcoxon test $P = 0.004$). Therefore, our machine-learning model "fraudulently" used the EoS to predict a random outcome.

Controlling the EoS when estimating the accuracy, the machine-learning model's MSE was 1.5, statistically significantly worse than the MSE of the null model (MSE=0.5, Wilcoxon test $P<0.001$). Our machine-learning could not use the MRI data to predict a random outcome, and indeed it only added noise.

5. Discussion

This manuscript first shows how to control the EoS when estimating accuracy - either conducting a separate study per site plus meta-analysis or including site as a covariate. We provide formulas for different accuracy measures and an R package to facilitate their application. Finally, we offer simulations that the reader may run and an example with real MRI data from internet-available repositories.

Interestingly, the simulations showed two unexpected results. First, they showed that the EoS might shrink (rather than inflate) the accuracy in some scenarios. We had initially considered this possibility at a theoretical level, and the simulations proved it in a specific situation: when there were very strong real effects in both sites. Second, the simulations showed that neither the covariate nor the meta-analysis methods perfectly control the EoS. That said, their biases were minor and substantially smaller than the biases observed when no controlling the EoS.

A reviewer of the current paper noted that another group had indeed reached very similar conclusions. Specifically, Dinga et al. (2020) reported that regressing out the confounding variables separately from each input variable before machine learning is insufficient, proposing a method to estimate accuracy similar to our covariate approach. The fact that two independent groups have noted a similar problem and have suggested addressing it with similar tools adds validity to their and our study. The papers also have several differences, e.g., we focus on the EoS because we think they are likely more complex effects and therefore more likely to be incompletely removed during the model's training. We may reasonably expect that differences between males and females, for example, may be primarily represented by a relatively simple additive factor in some voxels (males have a larger or smaller volume than females). Conversely, different MRI devices may create more complex EoS, such as creating different intricate spatial covariance patterns (e.g., adjacent voxels may be more correlated in one site than in another). Besides, the factor "site" may be on most occasions less representative than, for instance, the factor "sex." For the latter, our sample may easily include all groups present in the population. Thus, a prediction model that uses "sex" to predict may apply to the whole population. Conversely, our sample only includes a tiny subgroup of all potential sites. Thus, a prediction model that uses "site" to predict would apply to a tiny part of the population. That said, we acknowledge that it may be desirable to estimate the accuracy controlling variables other than the site in some circumstances. For these cases, we refer the reader to work by Dinga et al. Finally, we must acknowledge there is the possibility that the real effects and the EoS are collinear, which would prevent a clear differentiation of the real effects and the EoS.

To conclude, we hope that we have clearly shown the importance of controlling the EoS when estimating machine-learning studies' accuracy and how to do it. We fully acknowledge that it is theoretically possible that the method used to remove the EoS successfully removes all EoS in a specific dataset. However, we recommend not relying on this assumption because it may not hold. Indeed, there is no reason to rely on this assumption when it is easy to control the EoS with the R package we provide ("multisite.accuracy").

Declaration of Competing Interest

Dr. Vieta has received grants and served as consultant, advisor, or CME speaker for the following entities (work unrelated to the topic of this manuscript): AB-Biotics, Abbott, Allergan, Angelini, Dainippon Sumitomo Pharma, Galenica, Janssen, Lundbeck, Novartis, Otsuka, Sage, Sanofi-Aventis, and Takeda.

Acknowledgement

This work was supported by the Spanish Ministry of Science, Innovation and Universities / Economy and Competitiveness / Instituto de Salud Carlos III (CPII19/00009, PI19/00394, FI20/00047), co-financed by ERDF Funds from the European Commission ("A Way of Making Europe").

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.pscychresns.2021.111313.

References

Archer, D.B., Bricker, J.T., Chu, W.T., Burciu, R.G., McCracken, J.L., Lai, S., Coombes, S. A., Fang, R., Barmpoutis, A., Corcos, D.M., Kurani, A.S., Mitchell, T., Black, M.L., Herschel, E., Simuni, T., Parrish, T.B., Comella, C., Xie, T., Seppi, K., Bohnen, N.I., Muller, M., Albin, R.L., Krismer, F., Du, G., Lewis, M.M., Huang, X., Li, H., Pasternak, O., McFarland, N.R., Okun, M.S., Vaillancourt, D.E., 2019. Development and validation of the automated imaging differentiation in Parkinsonism (AID-P): a multisite machine learning study. The Lancet. Digital Health 1, e222–e231.

Beheshti, I., Demirel, H., Matsuda, H., Alzheimer's Disease Neuroimaging, I., 2017. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. Comput. Biol. Med. 83, 109–119.

Dinga, R., Schmaal, L., Penninx, B.W.J.H., Veltman, D.J., Marquand, A.F., 2020. Controlling for effects of confounding variables on machine learning predictions, bioRxiv.org.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.

Focke, N.K., Helms, G., Kaspar, S., Diederich, C., Toth, V., Dechent, P., Mohr, A., Paulus, W., 2011. Multi-site voxel-based morphometry–not quite there yet. Neuroimage 56, 1164–1170.

Fortin, J.P., Parker, D., Tunc, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multisite diffusion tensor imaging data. Neuroimage 161, 149–170.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33, 1–22.

Gill, S., Mouches, P., Hu, S., Rajashekar, D., MacMaster, F.P., Smith, E.E., Forkert, N.D., Ismail, Z., Alzheimer's Disease Neuroimaging, I., 2020. Using machine learning to predict dementia from neuropsychiatric symptom and neuroimaging data. J. Alzheimers Dis. 75, 277–288.

Glocker, B., Robinson, R., Castro, D.C., Dou, Q., Konukoglu, E., 2019. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects, arXiv.org.

Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21, 2409–2419.

Janes, H., Pepe, M.S., 2009. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika 96, 371–382.
Leger, C., Herbert, M., DeSouza, J.F.X., 2020. Non-motor clinical and biomarker predictors enable high cross-validated accuracy detection of early PD but lesser cross-validated accuracy detection of scans without evidence of dopaminergic deficit. Front Neurol 11, 364.
Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 19, 1498–1507.
Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., Senior, A. M., 2015. Meta-analysis of variation: ecological and evolutionary applications and beyond. Methods Ecol. Evol. 6, 143–152.
R Core Team, 2020. . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quide, Y., Green, M.J., Weickert, C.S., Weickert, T., Bruggemann, J., Kircher, T., Nenadic, I., Cairns, M.J., Seal, M., Schall, U., Henskens, F., Fullerton, J.M., Mowry, B., Pantelis, C., Lenroot, R., Cropley, V., Loughland, C., Scott, R., Wolf, D., Satterthwaite, T.D., Tan, Y., Sim, K., Piras, F., Spalletta, G., Banaj, N., Pomarol-Clotet, E., Solanes, A., Albajes-Eizagirre, A., Canales-Rodriguez, E.J., Sarro, S., Di Giorgio, A., Bertolino, A., Stablein, M., Oertel, V., Knochel, C., Borgwardt, S., du Plessis, S., Yun, J.Y., Kwon, J. S., Dannlowski, U., Hahn, T., Grotegerd, D., Alloza, C., Arango, C., Janssen, J., Diaz-Caneja, C., Jiang, W., Calhoun, V., Ehrlich, S., Yang, K., Cascella, N.G., Takayanagi, Y., Sawa, A., Tomyshev, A., Lebedeva, I., Kaleda, V., Kirschner, M., Hoschl, C., Tomecek, D., Skoch, A., van Amelsvoort, T., Bakker, G., James, A., Preda, A., Weideman, A., Stein, D.J., Howells, F., Uhlmann, A., Temmingh, H., Lopez-Jaramillo, C., Diaz-Zuluaga, A., Fortea, L., Martinez-Heras, E., Solana, E., Llufriu, S., Jahanshad, N., Thompson, P., Turner, J., van Erp, T., collaborators, E.C., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956.
Rodriguez-Alvarez, M.X., Roca-Pardinas, J., Cadarso-Suarez, C., 2011. ROC curve and covariates: extending induced methodology to the non-parametric framework. Stat Comput 21, 483–499.
Schwarzer, G., Chemaitelly, H., Abu-Raddad, L.J., Rucker, G., 2019. Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. Res Synth Methods 10, 476–483.

5.2. ARTICLE 2: COMBINING MRI AND CLINICAL DATA TO DETECT HIGH RELAPSE RISK AFTER THE FIRST EPISODE OF PSYCHOSIS

Schizophrenia

ARTICLE OPEN

Check for updates

Combining MRI and clinical data to detect high relapse risk after the first episode of psychosis

Aleix Solanes[1,2,3], Gisela Mezquida[1,4,5,6], Joost Janssen[5,7], Silvia Amoretti[1,4,5,8], Antonio Lobo[5,9,10], Ana González-Pinto[5,11,12,13], Celso Arango[5,7], Eduard Vieta[1,5,8,14], Josefina Castro-Fornieles[1,5,8,15], Daniel Bergé[3,5,16], Auria Albacete[2], Eloi Giné[17], Mara Parellada[5,7], Miguel Bernardo[1,4,5,8], PEPs group (collaborators)*, Edith Pomarol-Clotet[2,5,36]✉ and Joaquim Radua[1,2,5,18,19,36]✉

Detecting patients at high relapse risk after the first episode of psychosis (HRR-FEP) could help the clinician adjust the preventive treatment. To develop a tool to detect patients at HRR using their baseline clinical and structural MRI, we followed 227 patients with FEP for 18–24 months and applied MRIPredict. We previously optimized the MRI-based machine-learning parameters (combining unmodulated and modulated gray and white matter and using voxel-based ensemble) in two independent datasets. Patients estimated to be at HRR-FEP showed a substantially increased risk of relapse (hazard ratio = 4.58, $P < 0.05$). Accuracy was poorer when we only used clinical or MRI data. We thus show the potential of combining clinical and MRI data to detect which individuals are more likely to relapse, who may benefit from increased frequency of visits, and which are unlikely, who may be currently receiving unnecessary prophylactic treatments. We also provide an updated version of the MRIPredict software.

Schizophrenia (2022)8:100 ; https://doi.org/10.1038/s41537-022-00309-w

INTRODUCTION

The discovery of associations between magnetic resonance imaging (MRI) measures and mental disorders[1] led to an initial enthusiasm about finding MRI-based biomarkers, but we have failed so far. However, new machine-learning methods have reopened the possibility of creating MRI-based tools that, while far from perfect biomarkers, could still help the clinicians[2]. These tools could help the clinicians diagnose, predict the response to treatment, or estimate the risk of a bad outcome, adjusting the overall intervention accordingly.

Up to the moment, most MRI-based machine-learning studies have aimed to classify the individuals (e.g., patient vs. control, or between two diagnoses), and some other research has been devoted to creating models that estimate the risk of a bad outcome. For instance, many studies have investigated whether it is possible to use clinical data[3], MRI data[4], or their combination[5] to detect healthy individuals at high risk for psychosis. These studies have reported higher transition rates to psychosis in individuals that are males, have brief limited intermittent psychotic symptoms, or show reduced cortical gray matter[6,7].

Conversely, very little research has focused on detecting those patients with first episode of psychosis (FEP) at high relapse risk (HRR). This lack of research is striking because FEP represents one of the main challenges for mental health[8]. Without an appropriate differential diagnosis and early intervention, clinical development after FEP can lead to a chronic condition[9]. Detecting subjects at HRR is crucial since relapse puts their psychosocial recovery at risk, raises the chance of treatment resistance, and has been linked to higher direct and indirect social and economic costs[10]. A few studies have created models to estimate this risk based on clinical data[11,12], using variables such as the presence of manic and negative symptoms[13–15], the diagnosis[12,15], or cannabis use[11,16,17]. Fewer studies have created models to estimate the risk of outcomes other than relapse (e.g., the severity of future symptoms) based on brain MRI data[18,19], using volumetric brain changes during the first year[20] or voxel/surface-based data[18]. And to our knowledge, no studies have attempted to create MRI-based relapse risk-estimation models.

This lack of research is unfortunate, given that a structural MRI-based tool able to detect FEP-HRR would be clinically valuable and feasible. It would be valuable because even if the accuracy of the HRR-FEP detection was modest, it could help the clinician adjust the follow-up and treatment of the patients as deemed beneficial[21]. It would be feasible since individuals with a FEP may undergo an MRI to discard organic brain pathology, so that the structural MRI required for this tool would serve both. This better clinical management would reduce the number of relapse-related hospitalizations in patients at HRR-FEP and exclude patients at low relapse risk from therapies unnecessary for them. Therefore, it would improve the quality of life of individuals with a FEP and reduce the burden on National Health System expenditure.

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]FIDMAG Germanes Hospitalàries Research Foundation, Barcelona, Spain. [3]Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Barcelona, Spain. [4]Barcelona Clínic Schizophrenia Unit, Neuroscience Institute. Hospital Clínic de Barcelona, Barcelona, Spain. [5]Centro de Investigación Biomédica en Red en Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain. [6]Department of Clinical Foundations, Pharmacology Unit, University of Barcelona, Barcelona, Spain. [7]Department of Child and Adolescent Psychiatry, Institute of Psychiatry Mental Health, Hospital General Universitario Gregorio Marañón, School of Medicine, Universidad Complutense, IiSGM, Madrid, Spain. [8]Department of Medicine, University of Barcelona, Barcelona, Spain. [9]Department of Medicine and Psychiatry, Universidad de Zaragoza, Zaragoza, Spain. [10]Instituto de Investigación Sanitaria Aragón (IIS Aragón), Zaragoza, Spain. [11]Universidad del País Vasco / EHU, Leioa, Bizkaia, Spain. [12]Instituto de Investigación Sanitaria Bioaraba, Vitoria-Gasteiz, Alava, Spain. [13]Psychiatric Department, Hospital Universitario de Alava, Vitoria-Gasteiz, Alava, Spain. [14]Barcelona Clínic Bipolar and Depressive Disorders Unit, Institute of Neurosciences, Hospital Clínic de Barcelona, Barcelona, Spain. [15]Department of Child and Adolescent Psychiatry and Psychology, 2017SGR881. Institute of Neuroscience, Hospital Clínic, Barcelona, Spain. [16]Hospital del Mar Medical Research Institute, Barcelona, Spain. [17]Hospital de Mataró, Barcelona, Spain. [18]Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom. [19]Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden. [36]These authors jointly supervised this work: Edith Pomarol-Clotet, Joaquim Radua. *A list of authors and their affiliations appears at the end of the paper. ✉email: epomarol-clotet@fidmag.org; radua@recerca.clinic.cat

Published in partnership with the Schizophrenia International Research Society

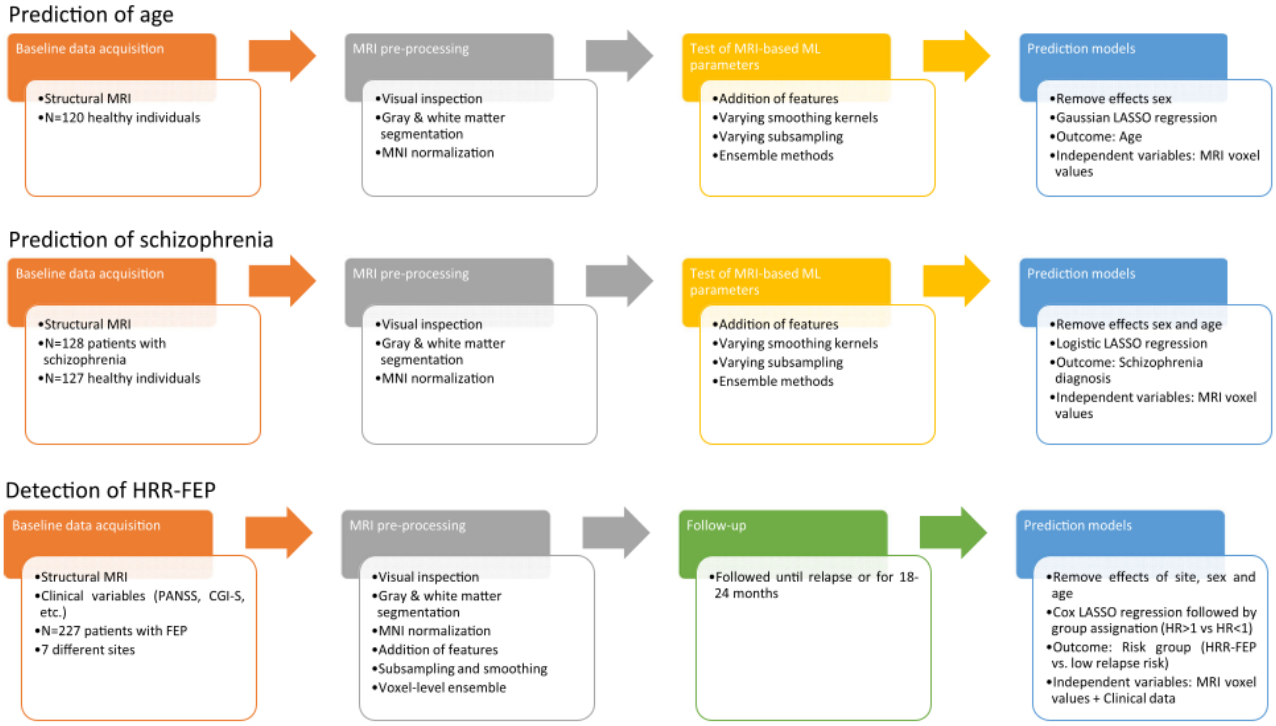


Fig. 1 Main study steps. Overall steps followed in this study.

The current study investigated whether structural MRI might help detect patients at HRR-FEP. To this end, we created an HRR-FEP detection tool. Additionally, we report how we previously optimized the MRI-based machine-learning parameters, using two independent datasets to avoid data leakage or over-complexity (see clarification later). We also freely provide the updated MRI-based machine-learning software to allow other groups to develop their own detection models and a website (see "Available resources") that estimates HRR-FEP quickly to help other groups independently replicate our model's accuracy assessment.

METHODS

See Fig. 1 for a view of the overall steps of the study. This study complies with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD, see checklist in the Supplement).

Participants

The cohort included 227 adolescents/adults with a FEP from 7 different hospitals in Spain, including a previous multicenter study[22,23], prospectively followed for two years. We invited all patients who met the inclusion criteria during the recruitment periods to join the study. We estimated the sample size based on a previous meta-analysis[24], in which the relapse rate at two years was around 37%. With this estimation, the overall sample size to detect a hazard ratio (HR) = 2 between patients at HRR-FEP and patients at low relapse risk had to be 190 according to R package powerSurvEpi (https://CRAN.R-project.org/package=powerSurvEpi). We included 20% more to compensate for potential early drop-outs. The mean age was 24.2 years (SD 7.4), and there were 78 females (34.4%) (Table 1). The sample included both young adolescents (12–14 years, $n = 6$) and late adolescents/adults (15–59, $n = 221$); as detailed later, to ensure that the estimation of the model accuracy is not confounded by mixing young adolescents with old adolescents/adults, we

Table 1. Description of the cohort ($n = 227$).

Age	24.2 (7.4)
Sex: female	78 (34.4%)
Familiar psychiatric history	110 (58.2%)
Affective*	54 (28.6%)
Suicide*	3 (1.4%)
Affective psychosis	48 (21.1%)
Baseline diagnosis	
Schizophrenia	78 (34.4%)
Bipolar disorder	41 (18.1%)
Schizoaffective disorder	12 (5.3%)
Substance-induced psychosis	10 (4.4%)
Major depressive disorder	5 (2.2%)
Other**	81 (35.7%)
Positive and Negative Syndrome Scale (PANSS)	71.2 (SD 24.4)
Positive scale	17.6 (SD 7.9)
Negative scale	18.2 (SD 8.5)
General psychopathology scale	35.9 (SD 12.7)
Global Assessment of Functioning scale (GAF)	50.5 (SD 19.7)
Clinical Global Impression scale (CGI-S)	4.4 (SD 1.1)
Young Mania Rating Scale (YMRS)	7.9 (SD 10.2)
Montgomery Asberg Depression rating scale (MADRS)	0.2 (SD 9.9)
Long-acting injectable antipsychotic	15 (6.6%)

*Familiar affective history included diagnoses such as bipolar disorder or MDD. Familiar suicide history included consummated attempts.
**Other baseline diagnoses included brief psychotic disorder, schizophreniform disorder, delusional disorder, and psychotic disorder not otherwise specified.
Data are presented as mean (SD) or number (%).

repeated the validation of the model after excluding young adolescents.

We defined relapses as exacerbations of symptoms during at least one week with at least one of eight PANSS items (P1, P2, P3, N1, N4, N6, G5, and G9) scoring above 3 (mild)[25]. On the contrary, remission was defined as scoring <3 in all eight PANSS items. We only considered relapse after at least 6 months of remission.

We detail the inclusion/exclusion criteria and a more detailed description of the cohort in the Supplement. The ethical committees of all hospitals had approved the study, conducted according to the Declaration of Helsinki. Furthermore, all participants and parents/legal guardians for adolescents under 16 had given written informed consent.

Collection and processing of baseline structural MRI data

We acquired a high-resolution structural image from each participant with a T1-weighted gradient-echo sequence with different devices (see Supplement for details). We used a voxel-based morphometry (VBM) pre-processing pipeline because we have previously found higher accuracy using VBM data[26] (see Supplement for details).

Removal of the effects of the site

The effects of the site (e.g., differences in MRI data due to using different devices) might increase noise and confound the analyses. To remove them, we used a recently developed method to control for batch effects named ComBat, as several studies have shown its superiority to simply adding "site" as a covariate in the linear models[27,28]. We found the ComBat parameters (i.e., the MRI differences between sites) using the processed images from exclusively the training set (i.e., we did not use the test set to find the parameters). We then removed the effects of the site from the processed images of both the training and the test sets using these parameters. We must highlight again that the effects of the site were estimated only using individuals from the training set (i.e., not a single piece of information from the test set), thus preventing any information leak. We have previously modified the ComBat functions to allow this separate estimation and application of the ComBat parameters[28].

We also controlled the effects of the site when estimating the model's accuracy (see details later), which is important because the effects of the site might bias the accuracy even when researchers attempted to remove them during the creation of the machine-learning model[29].

Optimization of MRI-based machine-learning parameters

We optimized the MRI-based machine-learning parameters using two independent datasets. Our main reason for using independent datasets was to avoid any data leakage. We reasoned that if we used the same cohort to optimize the machine-learning parameters and create the risk-estimation model, we could end up validating this model in patients we had previously used to optimize the parameters (based on the best relapse risk estimations). We acknowledge that one strategy to prevent such data leakage would be optimizing the parameters separately for each fold via within-fold cross-validation using the training sets exclusively. However, such a strategy could result in different MRI parameters for the different folds, creating over-complexity in the model. Rather, we looked for general MRI settings that would be stable not only for the different folds but for different predictions or studies.

One dataset included 120 healthy individuals[30], and we used their MRI data to predict a continuous variable (their age). The other dataset included 255 individuals, half of them with a schizophrenia diagnosis[26,30,31], and we used their MRI data to predict a binary variable (whether they had received the schizophrenia diagnosis or not). See the Supplement for details. The creation of machine-learning models was analog to the one described later.

We defined the default settings as unmodulated gray matter images, smoothed with a kernel of $\sigma = 4$ mm (corresponding to FWHM = 9.5 mm) and a voxel size of $3 \times 3 \times 3$ mm^3. We tested whether the accuracy of MRI-based machine-learning models depended on: the addition of features (gray and white matter images, modulated and unmodulated images as they convey complementary volumetric information[30], global gray matter volume and global brain volume, and the midline abnormalities cavum septum pellucidum and absence of adhesion interthalamica, previously reported as good predictors in FEP[31,32]), the size of the smoothing kernels (from $\sigma = 2$ to 6 mm, corresponding to FWHM $\approx$ 5.3–15.8 mm, i.e., encompassing the usual widths of standard neuroimaging software) since the previous literature differs in the optimal kernel size[18,26], or the use of ensemble methods. Ensemble learning methods seek better prediction performance and robustness by combining the predictions of different models. We used two ensemble methods: (a) we resampled the subjects with replacement 18 times and repeated the creation of the risk-estimation model with each of the 18 resampled datasets, and (b) we selected half of the brain 18 times (i.e., dividing the brain in different angles) and repeated the creation of the risk-estimation model with each of the 18 half brains. Any of these two ensemble methods resulted in 18 risk-estimation models, which we applied to the test set, resulting in 18 risk estimations per patient. Finally, we calculated the mean of the 18 risk estimations to obtain a single risk-estimation per patient.

On another note, we tested two approaches to reduce the computational cost: applying additional subsampling ($6 \times 6 \times 6$ mm^3 or $12 \times 12 \times 12$ mm^3, instead of $3 \times 3 \times 3$ mm^3) and limiting the analyses to statistically significant voxels ($P < 0.05$ uncorrected at the univariate analysis).

We defined the accuracy of the age predictions as the mean absolute error (MAE) between the predicted and the actual age and the accuracy of the diagnostic predictions as the proportion of correct predictions. Finally, we assessed whether differences in accuracy between the analysis using a given parameter and the reference analysis (unmodulated gray matter smoothed with $\sigma = 4$ mm) were statistically significant by conducting a paired-sample Wilcoxon test of the absolute errors of the two analyses.

Creation and validation of the HRR-FEP detection tool

We used a cross-validation scheme to create the tool using a set of patients and validate it using a new set of patients. Specifically, we randomly divided the overall cohort into ten groups or "folds" trying to preserve a similar number of relapses in each fold. First, we created the model using data from individuals from folds 2 to 10 (the "training set"), and we estimated the relapse risk of individuals from fold 1 (the "test set") (Fig. 2). We then created the model using data from individuals from folds 1 and 3–10, and we estimated the relapse risk of individuals from fold 2. And so on. Therefore, we could estimate the relapse risk of all individuals, but we never used the same individuals for training and validating the model.

The creation of the HRR-FEP models in the training set consisted of fitting a multiple regression. The dependent variable was the time to relapse. The independent variables were the clinical data (including the items from the symptom scales PANSS, GAF, MADRS, YMRS, the diagnosis, and whether the patient was taking long-acting injectable antipsychotic treatment) and the voxel values of the pre-processed MRI. Before conducting the regression, we removed the effects of age and sex from the training MRI data with standard linear models. We must highlight once more that the effects of age and sex were estimated only using individuals from the training set (i.e., not a single information from the test set), thus preventing any information leak. We also scaled

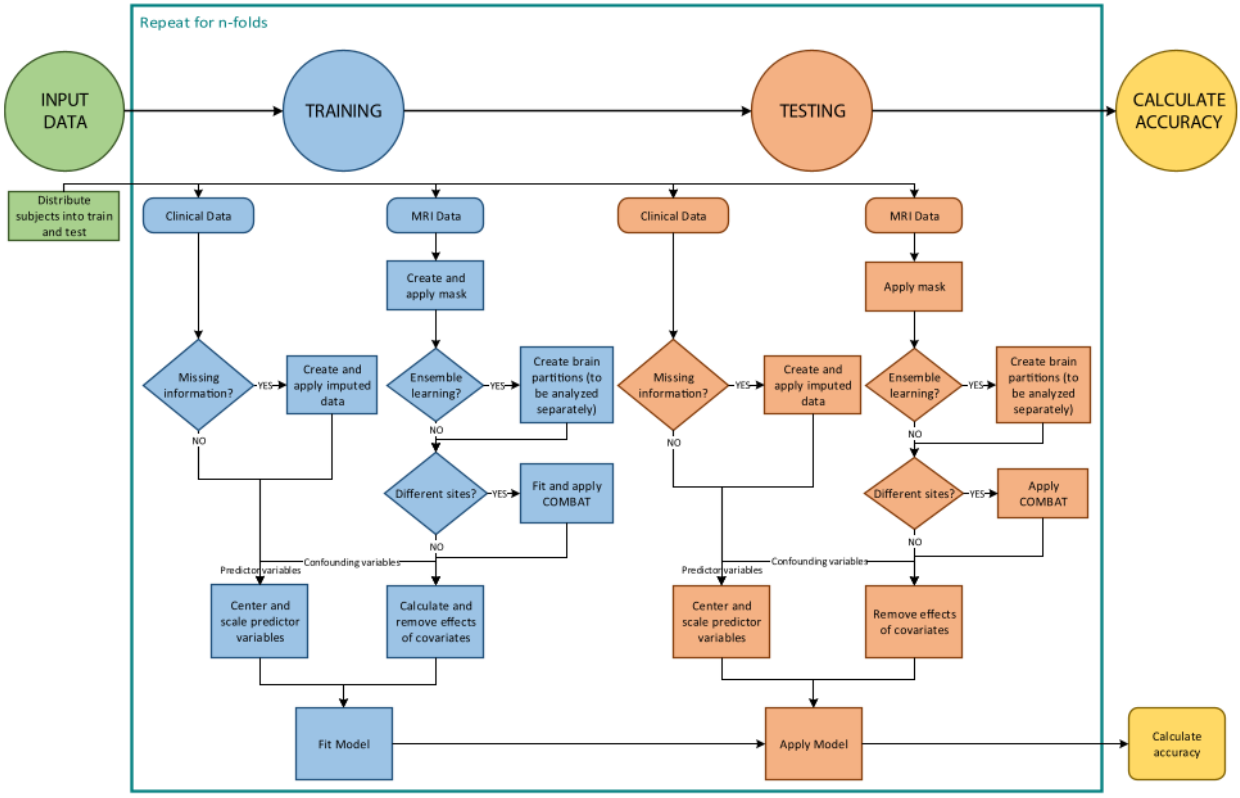


Fig. 2 MRIPredict flowchart. Creation of the high relapse risk after the first episode of psychosis (HRR-FEP) detection tool.

the clinical variables to the [0-1] range to have a distribution like the MRI voxels. To avoid overfitting, we used a "lasso" regression, which automatically selects a few regressors by penalizing the sum of the absolute value of the coefficients and has been proven to be able to deal with high-dimensional data and still achieve high-performance models[33]. A regularization parameter defines the amount of penalization, ranging from null (no penalization, as in a standard regression) to infinity (maximum penalization). This regularization parameter is automatedly selected by the algorithm via internal cross-validation within the training set. We chose the lasso regression algorithm for its good performance[26], simplicity, and adequacy for survival analyses. All these previous steps estimated using the training set were applied later to the test set to validate the performance of the model.

In other words, we found a risk-estimation model using the patients of the training set exclusively, and afterward, we applied the model to the patients of the test set to estimate their risk of relapse. To estimate a patient's risk, we multiplied each coefficient of the lasso model (see Table 3) by the value of the variable in the patient and added the results. If the sum was >0 (corresponding to a HR > 1), we considered that the patient was at HRR-FEP. Conversely, if the sum was ≤0 (corresponding to HR ≤ 1), we considered the patient at low relapse risk.

To test whether individuals estimated to be at HRR-FEP had statistically more relapses than individuals at low relapse risk, we used the "multisite.accuracy" package[29], which considers the site's residual effects when estimating the accuracy. Specifically, we conducted a mixed-effects Cox proportional hazards regression (https://CRAN.R-project.org/package=coxme). The dependent variable was the time to relapse. The independent variable was the estimated risk group (HRR-FEP vs. low relapse risk), and the site was a random factor of no interest.

To rule out whether the model's accuracy could mainly depend on MRI data or clinical data, we also created HRR-FEP detection tools exclusively based on MRI data or clinical data. Also, for descriptive purposes, we mapped the brain regions univariately associated with increased or decreased relapse risk after the FEP using standard survival analyses (see Supplement).

We conducted the analyses with our freely available graphical software MRIPredict (which we have updated for this work), based on the "glmnet" package for R (https://glmnet.stanford.edu/).

Available resources

Groups interested in conducting similar analyses can download our free graphical-user-interface MRIPredict software at https://www.mripredict.com/.

We encourage independent groups to replicate our model's accuracy assessment. To help them, we provide a website-based version of the tool (https://www.mripredict.com/hrr-fep/) that quickly estimates the HRR-FEP of an individual. For the website, we fitted a model using the whole cohort and selected the coefficients with an absolute value ≥0.05 (see Supplement); its risk estimations seem perfect (all relapses are in HRR-FEP individuals). However, this accuracy is inflated because it uses the same individuals for training and testing; we obtained a more reliable accuracy estimation with cross-validation (see next). In addition, we only offer this tool to support replication by other researchers; the tool estimations must be considered experimental.

RESULTS

Cohort description

There were 16 relapses, representing a 9.4% relapse rate at 24 months. Note that while the number of relapses was limited,

Table 2. Optimization of MRI-based machine-learning parameters.

Adjustment		Age predictions			Diagnostic predictions		
		MAE	Absolute *P* value[(a)]	Relative *P* value[(b)]	Accuracy	Absolute *P* value[(a)]	Relative *P* value[(b)]
Addition of features	No (reference)	7.4 years	<0.001	–	70.8%	<0.001	–
	+ white matter and modulated images	6.2 years	<0.001	0.006	68.6%	<0.001	0.039
	+ global volumes	7.4 years	<0.001	n.s.	70.8%	<0.001	n.s.
	+ midline abnormalities	–	–	–	70.7%	<0.001	n.s.
						<0.001	
Varying smoothing kernel width	σ = 2 mm	7.8 years	<0.001	n.s.	68.9%	<0.001	n.s.
	σ = 3 mm	7.2 years	<0.001	n.s.	69.4%	<0.001	n.s.
	σ = 4 mm (reference)	7.4 years	<0.001	–	70.8%	<0.001	–
	σ = 5 mm	7.5 years	<0.001	n.s.	71.2%	<0.001	n.s.
	σ = 6 mm	7.5 years	<0.001	n.s.	70.8%	<0.001	n.s.
						<0.001	
Subsampling	Single (reference)	7.4 years	<0.001	–	70.8%	<0.001	–
	Double subsampling	7.5 years	<0.001	n.s.	70.6%	<0.001	n.s.
	Triple subsampling	7.3 years	<0.001	n.s.	70.3%	<0.001	n.s.
	Only statistically significant	7.4 years	<0.001	n.s.	71.0%	<0.001	n.s.
						<0.001	
Ensemble	No (reference)	7.4 years	<0.001	–	70.8%	<0.001	–
	Subjects	8.5 years	<0.001	<0.001	64.4%	<0.001	0.001
	Voxels (half brains)	7.1 years	<0.001	0.002	73.2%	<0.001	0.034
						<0.001	
Optimal parameters	+ white matter and modulated images, triple subsampling, the ensemble of voxels	6.3 years	<0.001	<0.001	74.4%	<0.001	0.043

MAE mean absolute error.
(a) Wilcoxon test comparing the predictions obtained with these settings with the predictions obtained with a null model (i.e., predicting that all individuals have the average age of the sample for age predictions, or to flipping a coin for diagnostic predictions).
(b) Wilcoxon test comparing the predictions obtained with these setting with the predictions obtained with the reference settings (unmodulated gray matter smoothed with σ = 4 mm and no subsampling or ensemble).

it still yielded enough statistical power to detect meaningful differences in relapse risk between groups (e.g., using the R package powerSurvEpi, we estimated that we had 70%/80%/90% power to detect a HR = 4.3/5.7/9.5). The median time from scan to relapse (in patients who had a relapse during the follow-up) was 7.4 months, and the median time from scan to the last follow-up visit (in patients with no relapse during the follow-up) was 23.7 months. We detected no statistically significant differences in relapse risk between affective and non-affective psychosis or between diagnoses, except increased risk in patients with a schizoaffective disorder diagnosis (HR = 3.6, $P = 0.046$).

Optimal MRI-based machine-learning parameters

When optimizing the MRI-based machine-learning parameters, we found that adding gray and white matter images, unmodulated modulated images, and the use of a voxel-level ensemble improved the accuracy (Table 2). Conversely, using subject-level ensemble worsened the accuracy. The other varying parameters did not influence accuracy. We thus selected the addition of gray and white matter images, unmodulated and modulated images, and the use of a voxel-level ensemble for the HRR-FEP analyses. We also chose triple subsampling because it makes all calculations substantially less computationally expensive.

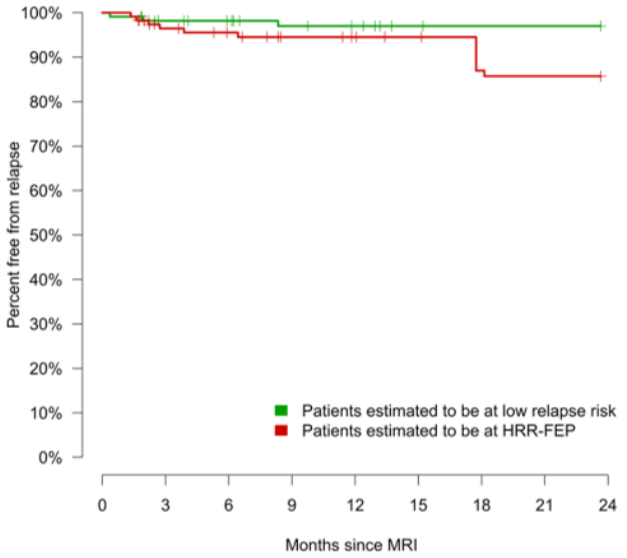


Fig. 3 Observed relapses depending on estimated risk group. Kaplan–Meier curves of the observed relapse in patients estimated to be at high relapse risk after the first episode of psychosis (HRR-FEP) vs. patients at low relapse risk.

Table 3. Descriptive univariate analysis and machine-learning estimators of high relapse risk after the first episode of psychosis (HRR-FEP).

	Descriptive univariate analysis	Machine learning
Clinical variables		
Schizoaffective disorder	HR = 3.6, $P = 0.046$	$\beta = +0.24$
↓ Poor rapport (PANSS N3)	–	$\beta = -0.01$
↓ Difficulty in abstract thinking (PANSS N5)	HR = 0.6, $P = 0.044$	$\beta = -0.074$
↓ Conceptual disorganization (PANSS P2)	–	$\beta = -0.01$
↓ Poor attention (PANSS G11)	–	$\beta = -0.04$
↑ Age in years	HR = 1.1, $P = 0.008$	–
↑ Long-acting injectable antipsychotic	–	$\beta = 0.01$
Gray matter increase		
↑ R Postcentral	–	Unm, [54, −6, 24], $\beta = +0.93$
Gray matter decrease		
↓ R middle temporal	–	Unm, [66, −6, −12], $\beta = -0.43$
↓ R inferior frontal/precentral	–	Unm, [30, 6, 36], $\beta = -0.21$ Mod, [42, 6, 36], $\beta = -0.18$
↓ R middle frontal	–	Unm, [30, 42, 36], $\beta = -0.20$
↓ R/L rectus	Unm, [−6, 30, −36], z = −2.6	Mod, [6, 30, −24], $\beta = -0.17$ Unm, [6, 30, −24], $\beta = -0.15$
↓ L superior frontal	Unm, [−18, 66, −24], z = −2.8	–
↓ R medial frontal	Unm, [6, 78, −12], z = −2.6	–
↓ R Angular	–	Unm, [30, −54, 36], $\beta = -0.05$
White matter increase		
↑ R precentral	–	Unm, [42, 6, 36], $\beta = +0.54$
↑ L Middle frontal	–	Unm, [−42, 6, 36], $\beta = +0.10$
White matter decrease		
↓ R middle frontal	–	Unm, [30, 30, 36], $\beta = -0.86$
↓ L inferior frontal	–	Mod, [−42, 18, 12], $\beta = -0.73$ Unm, [−42, 18, 12], $\beta = -0.57$
↓ R Cuneus	–	Mod, [18, −90, 12], $\beta = -0.18$
↓ R superior frontal	Unm, [6, 54, 24], z = 2.9	–
↓ L corpus callosum	Unm, [−18, 18, 24], z = 2.7	Unm, [−18, −30, 24], $\beta = -0.05$
↓ R corpus callosum	–	Mod, [6, 30, 0], $\beta = -0.05$
↓ L Middle frontal	–	Mod, [−30, 42, 12], $\beta = -0.06$
↓ R postcentral	–	Mod, [54, −6, 24], $\beta = -0.09$

L left, *Mod* modulate, *PANSS* Positive and Negative Syndrome Scale, *R* right, *Unm* unmodulated.
In the descriptive univariate analysis, we only report the peaks of MRI clusters with voxel uncorrected *P* value <0.005 and the clinical variables with uncorrected *P* value <0.05. For the sake of simplicity, we only report here the machine-learning coefficients with an absolute value ≥0.01 for clinical variables and ≥0.05 for MRI voxels; see the entire model in the Supplement.

HRR-FEP detection

The Cox regression of the time to relapse comparing patients estimated to be at HRR-FEP vs. low relapse risk was clinically relevant (HR = 4.58, i.e., HRR-FEP patients had five times more risk to relapse) and had a (borderline) statistical significance (HR 95% confidence interval = 1.01–20.74, Z = 1.98, $P = 0.048$, Fig. 3). The results were identical when we excluded young adolescents (i.e., 12–14-years-old). In the 114 individuals estimated to be at HRR-FEP, there were 13 relapses, representing a 14.8% relapse rate at 24 months. Conversely, there were only three relapses in the 113 individuals estimated to not be at HRR-FEP, representing a 2.9% relapse rate at 24 months. Using the R package powerSurvEpi, we estimated that the power to detect a HR = 4.58 with 16 relapses is 72%. The variables automatedly selected by the lasso regression to create the HRR-FEP detection tool were the diagnosis of schizoaffective disorder, the lack of difficulty in abstract thinking and poor impulse control, and the increase or decrease of unmodulated and modulated gray and white matter in several brain regions. Table 3 details the specific brain regions and clinical variables detected in the descriptive univariate analysis and the machine-learning model. We report the entire machine-learning model in the Supplement.

The HRR-FEP detection tools exclusively using MRI data or solely based on clinical variables failed to detect patients at HRR-FEP.

DISCUSSION

In this work, we created an MRI-based machine-learning tool to detect those patients at HRR-FEP using a cohort of 227 individuals with a FEP. The model showed to detect HRR-FEP successfully. The hazard of relapse was 4.5 larger in individuals estimated to be at HRR-FEP than in low relapse risk individuals (14.8% vs. 2.9% relapse rate at 2 years), and we estimated the power to detect such a hazard ratio of 4.5 with 16 relapses is 72%.

The study thus achieved the aim of creating a tool that may provide valuable information to the mental health professional. Ideally, the clinician could input the tool with a few MRI and clinical data to know if the patient is estimated to be at HRR-FEP or not,

and thus adjust the prophylactic treatment. Knowing this information early is important because currently, clinicians can only know which patients are at HRR-FEP after several relapses. And before that, patients at HRR-FEP may experience repeated relapses if the prevention is too weak, while patients at low relapse risk may experience increased adverse events if the prevention is too strong. That said, any adjustment of the prophylactic treatment should follow the "first, do no harm" principle because in our cohort, most (85%) individuals estimated to be at HRR-FEP did indeed not relapse. Not less important, the clinician could also consider removing or reducing the prophylactic treatment in individuals estimated to be unlikely to relapse. These patients currently may be receiving treatments that, if the patient is truly unlikely to relapse, may be little useful while harmful.

However, in any case, we want to highlight the need to validate the HRR-FEP detection tool before recommending it. We have noted previously that independent studies often fail to replicate the accuracy reported in mental health machine-learning publications[34], and our study may not be an exception. We cannot share participant data for privacy reasons. However, we provide the trained classifiers online so that independent researchers can still try to replicate our study results. This approach has been stated to be one of the most convincing forms of replication[35]. However, without intending to create hype, we also think that our work shows the potential clinical utility of MRI-based machine-learning when understood as a source of additional information for the psychiatrist.

We also want to highlight that this tool could be complemented by other tools that update the relapse risk during the follow-up. For example, we have reported for other disorders that the relapse risk at 12 months substantially decreases in patients who have been relapse-free for at least one year[24]. Thus, some patients initially at HRR-FEP may later be at low relapse risk. Similarly, information about changes in the first months could also likely offer valuable information for updating the risk estimation[20]. In this context, we would like to note that, as far as relapses also depend on events that will happen during the follow-up, it is unlikely that a machine-learning model that only uses baseline data scan achieves high risk-estimation accuracy.

A particularity of our study is that, instead of focusing on detecting those healthy individuals at high risk for FEP, it focuses on detecting those FEP patients at HRR. Many studies have already been published regarding predicting transition to psychosis, with varied results[3,4]. Conversely, no studies have been conducted to estimate HRR-FEP from MRI data to our knowledge. This lack of research is striking because assessing the relapse risk is essential to properly adjusting the preventive antipsychotic dose.

Our tool requires an MRI, but patients with a FEP may indeed already undergo an MRI to discard organic brain pathology, so that the structural MRI required for our tool would serve both. This fact increases the feasibility of the HRR-FEP detection tool, given that for many patients, it would only involve minor calculations on any computer. The context is different, for instance, for the detection of individuals with a higher risk of psychosis in the general population, where screening detection tools should only require inputting a small amount of available information. An example of such a screening detection tool is the Psychosis Polyrisc Score (PSS)[36], which only asks about the presence of a few risk factors[37] and has shown feasible in a real-world digital implementation[38].

Interestingly, the accuracy of HRR-FEP detection tools was poorer when we created machine-learning models that used only clinical data or only MRI data. Ad hoc, it may seem evident that the more information, the better the detection. However, many previous studies only used MRI to find biomarkers that should surpass clinical judgment. These may include serum component protein 4 (C4)[39], polygenic related Risk Score (PRS)[40], neuroanatomical variables[18,20]. Thus, poetically, we have found that, in the fight between clinical-based and biomarker-based psychiatry, joining efforts predicts better.

One key variable selected by the lasso regression was the diagnosis of schizoaffective disorder; this partly agrees with previous studies reporting associations of diagnosis or manic symptoms with increased relapse rate[12,13,15]. In addition, we think that in the current debate about the validity of DSM/ICD diagnoses, it is worth noting that diagnostic labels more than clinical scales helped predict future relapses. That said, this debate is entirely out of the scope of this paper. On another note, the protective effects of the difficulty in abstract thinking and poor impulse control are intriguing. We speculate that these symptoms may be related to latent disorder subtypes that might be clearer in subsequent phases of the illness. Finally, we must acknowledge that the variables showing statistical significance in the descriptive univariate analysis (see Supplement) were primarily different from the variables selected by the lasso regression. However, this disagreement is expectable because the latter only aims to predict and thus discards brain regions that do not add much to the prediction accuracy, even if they are statistically significant when considered alone.

Before creating the HRR-FEP detection tool, we used two independent datasets to find the optimal parameters for VBM-based machine learning. Finding the optimal parameters in two different datasets keeps the main study data unseen until we create the model for the HRR -FEP detection tool. We acknowledge that the accuracy of the age predictions was lower than that reported elsewhere[41]. This lower accuracy was probably related to the limited sample size of the age prediction dataset. However, we only aimed to compare the accuracy depending on different parameters. We found that the optimal parameters were the addition of gray and white matter images, the addition of unmodulated and modulated images, and the use of voxel-level ensemble. We encourage future studies to use these parameters. Also, we found that even triple subsampling did not affect the accuracy while substantially reducing computational costs.

We want to comment that, while previous work has searched for gold biomarkers with little success, this work shows the potential clinical use of MRI-based machine learning in risk assessment. We speculate that such risk assessment will very likely be far from perfect, i.e., we will not be able to know for sure which patients will have a relapse and which will not, or the date of the relapse. Indeed, such predictions may seem unrealistic considering that relapses also depend on life events and stressors after the assessment[42]. However, the estimation will be clinically valuable as far as we can estimate risk with enough accuracy to help the physician, i.e., so that the information translates into in an effective improvement of the care. Our study does not provide this level of accuracy yet, but we hope to have made a step for future studies.

This work has some limitations. First, this sample does not include the patients who did not meet the inclusion criteria or refused to participate in the study, who may differ from those included. It is a common limitation in many other studies. Second, even if we included 227 patients and followed them for 18–24 months, representing one of the largest brain imaging FEP cohorts worldwide, there were only 16 relapses. This relapse rate is lower than those reported in some previous cohorts[24,43,44]. To check whether the difference in relapse rate was due to our relapse criteria being only based on PANSS while others also considered hospitalizations, we retrieved hospitalizations, and the updated relapse rate (37%) was more in agreement with previous cohorts. However, we could not successfully repeat the analyses with hospitalizations because this information was unavailable on some sites. Third, the statistical significance was weak, probably due to our cohort's limited number of relapses. In any case, the power to detect a hazard of relapse of 4.5 with the sample size and the number of relapses in this study was 72%, very close to the conventional 80% required in sample size calculations. Fourth, more complex machine-learning algorithms, such as neural networks, might detect more patterns than the relatively simple

algorithms used here. However, these algorithms usually require substantially larger cohorts, which may be challenging to achieve. Fifth, for simplicity, we considered patients estimated to have a HR > 1 at HRR-FEP. However, the optimal division between groups could be at another HR threshold. Future studies evaluating the benefits and costs of the interventions at different HR levels may provide more insights into this question. Sixth, we could not evaluate medication adherence, DUP, and premorbid functioning because data was missing in some sites. Due to its established role in relapse, the use of these variables could improve model accuracy. Finally, we could not report statistics such as sensitivity and specificity. We could not estimate such statistics because our data was not binary (relapse vs. not relapse). Note that 38% of patients did not complete the follow-up, and thus we could not classify them as relapse or not relapse - we knew that they had not relapsed until the last visit, but we did not know if they had relapsed afterward. However, even if there were no follow-up losses, we would still report the Cox regression as the primary validation statistic because it considers whether relapses occurred earlier or later. In contrast, binary statistics do not.

To conclude, this study might represent a step towards a translational application of neuroimaging to mental health. Up to now, brain imaging prediction models have mainly aimed to imitate clinical judgment, for example, by training a support vector machine to differentiate between patients and controls based on their brain images[45]. Conversely, we combined clinical and MRI data to improve the accuracy of a tool that, instead of finding reliable biomarkers, aims to help the clinician, ultimately paving the way toward more personalized medicine in mental disorders.

DATA AVAILABILITY

Data are available upon request to the Research Ethics Committees of Benito Menni CASM, Hospital General de Granollers, Hospital de Mataró, Hospital Sant Rafael, Hospital de Bellvitge, Hospital Clínic de Barcelona, Hospital Universitario 12 de Octubre, Hospital Clínic de València, Hospital del Mar, Instituto de Investigación Sanitaria Aragón, Hospital General Universitario Gregorio Marañón, Hospital Sant Joan de Déu Barcelona, Hospital Santiago Apóstol de Vitoria-Gasteiz, and Servicio de Salud del Principado de Asturias.

CODE AVAILABILITY

Groups interested in conducting similar analyses can download and check the R code of the software used to conduct the analysis of this study, and use our free graphical-user-interface MRIPredict software at https://www.mripredict.com/.

Received: 10 June 2022; Accepted: 28 October 2022;
Published online: 17 November 2022

REFERENCES

1. DeLisi, L. E. et al. Cerebral ventricular enlargement as a possible genetic marker for schizophrenia. *Psychopharmacol. Bull.* **21**, 365–367 (1985).
2. Radua, J. & Carvalho, A. F. Route map for machine learning in psychiatry: absence of bias, reproducibility, and utility. *Eur. Neuropsychopharmacol.* **50**, 115–117 (2021).
3. Rosen, M. et al. Towards clinical application of prediction models for transition to psychosis: a systematic review and external validation study in the PRONIA sample. *Neurosci. Biobehav. Rev.* **125**, 478–492 (2021).
4. Smieskova, R. et al. Neuroimaging predictors of transition to psychosis-a systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **34**, 1207–1222 (2010).
5. Schmidt, A. et al. Improving prognostic accuracy in subjects at clinical high risk for psychosis: systematic review of predictive models and meta-analytical sequential testing simulation. *Schizophrenia Bull.* **43**, 375–388 (2017).
6. Salazar de Pablo, G. et al. Probability of transition to psychosis in individuals at clinical high risk: an updated meta-analysis. *JAMA Psychiatry* https://doi.org/10.1001/jamapsychiatry.2021.0830 (2021).
7. Fortea, A. et al. Cortical gray matter reduction precedes transition to psychosis in individuals at clinical high-risk for psychosis: a voxel-based meta-analysis. *Schizophrenia Res.* **232**, 98–106 (2021).
8. Harrison, G. et al. Recovery from psychotic illness: a 15- and 25-year international follow-up study. *Br. J. Psychiatry* **178**, 506–517 (2001).
9. Bernardo, M. et al. The prevention of relapses in first episodes of schizophrenia: the 2EPs Project, background, rationale and study design. *Revista de psiquiatria y salud mental* **14**, 164–176 (2021).
10. Ascher-Svanum, H. et al. The cost of relapse and the predictors of relapse in the treatment of schizophrenia. *BMC Psychiatry* **10**, 2 (2010).
11. Bhattacharyya, S. et al. Individualized prediction of 2-year risk of relapse as indexed by psychiatric hospitalization following psychosis onset: model development in two first episode samples. *Schizophrenia Res.* **228**, 483–492 (2021).
12. Puntis, S., Whiting, D., Pappa, S. & Lennox, B. Development and external validation of an admission risk prediction model after treatment from early intervention in psychosis services. *Transl. Psychiatry* **11**, 35 (2021).
13. Arrasate, M. et al. Prognostic value of affective symptoms in first-admission psychotic patients. *Int. J. Mol. Sci.* **17**, https://doi.org/10.3390/ijms17071039 (2016).
14. Wunderink, L. et al. Negative symptoms predict high relapse rates and both predict less favorable functional outcome in first episode psychosis, independent of treatment strategy. *Schizophrenia Res.* **216**, 192–199 (2020).
15. Hui, C. L. et al. Predicting first-episode psychosis patients who will never relapse over 10 years. *Psychological Med.* **49**, 2206–2214 (2019).
16. Berge, D. et al. Predictors of relapse and functioning in first-episode psychosis: a two-year follow-up study. *Psychiatric Services* **67**, 227–233 (2016).
17. Schoeler, T. et al. Poor medication adherence and risk of relapse associated with continued cannabis use in patients with first-episode psychosis: a prospective analysis. *Lancet. Psychiatry* **4**, 627–633 (2017).
18. Nieuwenhuis, M. et al. Multi-center MRI prediction models: predicting sex and illness course in first episode psychosis patients. *NeuroImage* **145**, 246–253 (2017).
19. Dazzan, P. et al. Clinical utility of MRI scanning in first episode psychosis. *Schizophrenia Bull.* **44**, S50–S51 (2018).
20. Cahn, W. et al. Brain volume changes in the first year of illness and 5-year outcome of schizophrenia. *Br. J. Psychiatry* **189**, 381–382 (2006).
21. Alvarez-Jimenez, M., Parker, A. G., Hetrick, S. E., McGorry, P. D. & Gleeson, J. F. Preventing the second episode: a systematic review and meta-analysis of psychosocial and pharmacological trials in first-episode psychosis. *Schizophrenia Bull.* **37**, 619–630 (2011).
22. Pina-Camacho, L. et al. Age at first episode modulates diagnosis-related structural brain abnormalities in psychosis. *Schizophrenia Bull.* **42**, 344–357 (2016).
23. Berge, D. et al. Elevated extracellular free-water in a multicentric first-episode psychosis sample, decrease during the first 2 years of illness. *Schizophrenia Bull.* https://doi.org/10.1093/schbul/sbz132 (2020).
24. Radua, J., Grunze, H. & Amann, B. L. Meta-analysis of the risk of subsequent mood episodes in bipolar disorder. *Psychother. Psychosomatics* **86**, 90–98 (2017).
25. Andreasen, N. C. et al. Remission in schizophrenia: proposed criteria and rationale for consensus. *Am J Psychiatry* **162**, 441–449 (2005).
26. Salvador, R. et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS ONE* **12**, e0175683 (2017).
27. Fortin, J. P. et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167**, 104–120 (2018).
28. Radua, J. et al. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage* **218**, 116956 (2020).
29. Solanes, A. et al. Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. *Psychiatry Res. Neuroimaging* **314**, 111313 (2021).
30. Radua, J. et al. Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies. *Front. Psychiatry* **5**, 13 (2014).
31. Landin-Romero, R. et al. Midline brain abnormalities across psychotic and mood disorders. *Schizophrenia Bull.* **42**, 229–238 (2016).
32. Kasai, K. et al. Cavum septi pellucidi in first-episode schizophrenia and first-episode affective psychosis: an MRI study. *Schizophrenia Res.* **71**, 65–76 (2004).
33. Greenshtein, E. & Ritov, Y. A. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988, 918 (2004).
34. Radua, J. What is the actual accuracy of clinical prediction models? The case of transition to psychosis. *Neurosci. Biobehavioral Rev.* **127**, 502–503 (2021).
35. Young, J., Kempton, M. J. & McGuire, P. Using machine learning to predict outcomes in psychosis. *Lancet Psychiatry* **3**, 908–909 (2016).
36. Oliver, D., Radua, J., Reichenberg, A., Uher, R. & Fusar-Poli, P. Psychosis polyrisk score (PPS) for the detection of individuals at-risk and the prediction of their outcomes. *Front. Psychiatry* **10**, 174 (2019).
37. Radua, J. et al. What causes psychosis? An umbrella review of risk and protective factors. *World Psychiatry* **17**, 49–66 (2018).
38. Oliver, D. et al. Real-world digital implementation of the psychosis polyrisk score (PPS): a pilot feasibility study. *Schizophrenia Res.* **226**, 176–183 (2020).
39. Mondelli, V. et al. Baseline high levels of complement component 4 predict worse clinical outcome at 1-year follow-up in first-episode psychosis. *Brain Behav. Immunity* **88**, 913–915 (2020).

40. Harrisberger, F. et al. Impact of polygenic schizophrenia-related risk and hippocampal volumes on the onset of psychosis. *Transl. Psychiatry* **6**, e868–e868 (2016).
41. Baecker, L. et al. Brain age prediction: a comparison between machine learning models using region- and voxel-based morphometric data. *Human Brain Mapping* **42**, 2332–2346 (2021).
42. Simhandl, C., Radua, J., Konig, B. & Amann, B. L. The prevalence and effect of life events in 222 bipolar I and II patients: a prospective, naturalistic 4 year follow-up study. *J. Affective Disorders* **170**, 166–171 (2015).
43. Robinson, D. et al. Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. *Archives General Psychiatry* **56**, 241–247 (1999).
44. Tiihonen, J. et al. A nationwide cohort study of oral and depot antipsychotics after first hospitalization for schizophrenia. *Am. J. Psychiatry* **168**, 603–609 (2011).
45. Nieuwenhuis, M. et al. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *NeuroImage* **61**, 606–612 (2012).

ACKNOWLEDGEMENTS

We are grateful to all participants. We would also like to thank the Instituto de Salud Carlos III, the Spanish Ministry of Science, Innovation, and Universities, the European Regional Development Fund (ERDF/FEDER), European Social Fund, "Investing in your future", "A way of making Europe" (projects: PI08/0208, PI11/00325, PI14/00292, PI14/00612, PI14/01148, PI14/01151, PI17/00481, PI17/01997, PI18/01055, PI19/00394, PI19/00766, PI20/00721, PI20/01342, and PI21/00713; contracts: CP14/00041, JR19/00024, CPII19/00009, and CD20/00177); CERCA Program; Catalan Government, the Secretariat of Universities and Research of the Department of Enterprise and Knowledge (2017SGR01271, 2017SGR1355, and 2017SGR1365); Institut de Neurociències, Universitat de Barcelona; Madrid Regional Government (B2017/BMD-3740 AGES-CM-2); the University of the Basque Country (2019 321218ELCY GIC18/107); the Basque Government (2017111104); European Union Structural Funds, European Union Seventh Framework Program, European Union H2020 Program under the Innovative Medicines Initiative 2 Joint Undertaking (grant agreement No 115916, Project PRISM, and grant agreement No 777394, Project AIMS-2-TRIALS), Fundación Familia Alonso, Fundación Alicia Koplowitz and Fundación Mutua Madrileña. The funding organizations played no role in the study design, data collection, analysis, or manuscript approval.

AUTHOR CONTRIBUTIONS

J.R. and A.S. contributed to the data analysis and writing of the paper. J.R. and E.P. designed the study. E.P., J.R., G.M., J.J., S.A., A.L., A.G., C.A., E.V., J.C., D.B., A.A., E.G., M.P., M.B., and PEPs group (collaborators) contributed in the data collection. J.R. contributed acquiring funding. All authors critically revised the manuscript and approved the completed version. All authors are accountable for this work.

COMPETING INTERESTS

Dr. Bernardo has been a consultant for, received grant/research support and honoraria from, and been on the speakers/advisory board of AB-Biotics, Adamed, Angelini, Casen-Recordati, Janssen-Cilag, Menarini, Rovi, and Takeda. Dr. C. De-la-Camara received financial support to attend scientific meetings from Janssen, Almirall, Lilly, Lundbeck, Rovi, Esteve, Novartis, AstraZeneca, Pfizer, and Casen-Recordati. CDC has received honoraria from Sanofi and Exeltis. Dr. Parellada has received educational honoraria from Otsuka, research grants from Instituto de Salud Carlos III (ISCIII), Ministry of Health, Madrid, Spain, has received grant support from ISCIII, Horizon2020 of the European Union, CIBERSAM, Fundación Alicia Koplowitz, and Mutua Madrileña and travel grants from Otsuka, Exeltis and Janssen; she has served as a consultant for Servier, Exeltis, Fundación Alicia Koplowitz, and ISCIII. LPC has received honoraria or grants unrelated to the present work from Rubió, Rovi, and Janssen. Dr. R. Rodriguez-Jimenez has been a consultant for, spoken in activities of, or received grants from Instituto de Salud Carlos III, Fondo de Investigación Sanitaria (FIS), Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid Regional Government (S2010/BMD-2422 AGES; S2017/BMD-3740), Janssen-Cilag, Lundbeck, Otsuka, Pfizer, Ferrer, Juste, Takeda, Exeltis, Casen-Recordati, Angelini. Dr. Arango has been a consultant to or has received honoraria or grants from Acadia, Angelini, Biogen, Boehringer, Gedeon Richter, Janssen Cilag, Lundbeck, Medscape, Menarini, Minerva, Otsuka, Pfizer, Roche, Sage, Servier, Shire, Schering Plough, Sumitomo Dainippon Pharma, Sunovion and Takeda. Dr. Vieta has received grants and served as a consultant, advisor, or CME speaker for the following entities (work unrelated to the topic of this manuscript): AB-Biotics, Abbott, Allergan, Angelini, Dainippon Sumitomo Pharma, Galenica, Janssen, Lundbeck, Novartis, Otsuka, Sage, Sanofi-Aventis, and Takeda. The remaining authors report no financial relationships with commercial interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41537-022-00309-w.

Correspondence and requests for materials should be addressed to Edith Pomarol-Clotet or Joaquim Radua.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PEPs GROUP (COLLABORATORS)

Miquel Bioque[1,4,5,8], Constanza Morén[20,21], Laura Pina-Camacho[5,7], Covadonga M. Díaz-Caneja[5,7], Iñaki Zorrilla[5,11,12,13], Edurne Garcia Corres[5,11,12,13], Concepción De-la-Camara[5,9,10,22], Fe Barcones[5,9,10,23], María José Escarti[5,24,25], Eduardo Jesus Aguilar[5,24,25,26], Teresa Legido[16], Marta Martin[16], Norma Verdolini[1,5,8,14], Anabel Martinez-Aran[1,5,8,14], Immaculada Baeza[5,15], Elena de la Serna[1,5,8,15], Fernando Contreras[5,27], Julio Bobes[5,28,29,30,31], María Paz García-Portilla[5,28,29,30,31], Luis Sanchez-Pastor[32], Roberto Rodriguez-Jimenez[5,32,33], Judith Usall[34], Anna Butjosa[5,35], Pilar Salgado-Pineda[2,5] and Raymond Salvador[2,5]

[20]Centro de Investigación Biomédica en Red (CIBER) de Enfermedades Raras (CIBERER), Madrid, Spain. [21]Cellex, IDIBAPS, University of Barcelona-Hospital Clínic of Barcelona, Barcelona, Spain. [22]Hospital Clínico Universitario, Zaragoza, Spain. [23]Servicio Aragonés de la Salud, Centro de Salud de Tarazona, Zaragoza, Spain. [24]Department of Psychiatry, Hospital Clínico Universitario de Valencia, Valencia, Spain. [25]Biomedical Research Institute INCLIVA, Valencia, Spain. [26]Department Psychiatry, Faculty of Medicine, University of Valencia, Valencia, Spain. [27]Psychiatry Unit. Bellvitge University Hospital. IDIBELL, Barcelona, Spain. [28]Department of Psychiatry, Universidad de Oviedo, Oviedo, Spain. [29]Servicio de Salud del Principado de Asturias (SESPA), Oviedo, Spain. [30]Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Oviedo, Spain. [31]Instituto Universitario de Neurociencias del Principado de Asturias (INEUROPA), Oviedo, Spain. [32]Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), Madrid, Spain. [33]CogPsy Group, Universidad Complutense de Madrid (UCM), Madrid, Spain. [34]Parc Sanitari Sant Joan de Déu, Teaching, Research & Innovation Unit, Institut de Recerca Sant Joan de Déu, Barcelona, Spain. [35]Hospital Infanto-juvenil Sant Joan de Déu, Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain.

6. REPTES I FUTURES VIES DE TREBALL

En aquesta tesi he estat treballant en un mètode per tal de poder estimar el risc de recaure en pacients amb un primer episodi psicòtic mitjançant certes dades clíniques i imatges de RM. A més també he parlat de mètodes per tal de poder manejar dades multicèntriques en aquest tipus d'estudis, i d'aquesta forma possibilitar aconseguir mostres més grans, un dels grans requisits per millorar els models actuals.

Vist això, en aquesta secció introduiré algunes de les vies de futur treball per tal de poder seguir millorant en els estudis d'aprenentatge automàtic per tal de permetre'n algun dia la seva translació clínica.

6.1. ESTUDIS LONGITUDINALS

Els mètodes basats en dades d'un sol punt de temps poden ser útils. Tot i això, els canvis amb el pas del temps poden proporcionar informació rellevant per crear models del que pot passar (per exemple, si el pacient respondrà al tractament o tindrà alguna complicació). Per exemple, se sap que els pacients amb un primer episodi de psicosi mostren una disminució en el temps de la substància grisa cortical en comparació amb els controls sans [108], o que la reducció progressiva del volum de substància grisa en el gir temporal superior s'associa a una baixa millora dels símptomes de psicosi positiva [109]. Tenir un conjunt de dades recollides de centenars o milers de persones amb condicions similars durant un període prolongat permetrà trobar patrons més complexos. Aquests patrons permetran una millor predicció de resultats futurs. Per tant, els estudis longitudinals seran crucials per millorar la fiabilitat i el rendiment de les eines de suport a la presa de decisions en salut mental.

6.2. MOSTRES MÉS GRANS

Un dels primers passos habituals a l'hora de preprocessar dades de neuroimatge és la reducció de la dimensionalitat, mitjançant la selecció de característiques dissenyada per experts o l'extracció de característiques. Aquest procés augmenta el rendiment dels algorismes, però elimina la informació de les dades d'entrada. Per contra, algorismes

moderns com l'aprenentatge profund poden utilitzar dades d'entrada mínimament preprocessades i aprofitar els patrons subtils que normalment es retiren durant el preprocessament [110]. No obstant això, tot i que ja s'usa en algunes tasques de detecció d'anomalies cerebrals, l'aprenentatge profund encara no s'ha aplicat àmpliament per detectar subjectes precoços amb risc de desenvolupar un trastorn o un resultat rellevant. Una raó crítica per no fer servir algorismes d'aprenentatge profund és que requereixen, en general, conjunts de dades substancialment més grans que altres enfocaments d'aprenentatge automàtic.

Els conjunts de dades de neuroimatge solen ser difícils d'adquirir. Així i tot, consorcis emergents, com el consorci ENIGMA (http://enigma.ini.usc.edu/), ja estan permetent fer anàlisis sobre grans conjunts de dades d'una altra manera impossibles de reclutar [91]. De fet, una mostra multisite més gran no només millora la potència estadística dels estudis i permet l'ús de l'aprenentatge profund, sinó que també millora la generalitzabilitat dels models a noves dades.

6.3. NOUS ALGORISMES

Els algoritmes i mètodes evolucionen cada dia, així que potser la millor eina per detectar subjectes en risc encara està per aparèixer. En aquesta secció, només ratllarem la superfície i revisarem alguns dels mètodes més prometedors en aprenentatge automàtic.

6.3.1. ALGORISMES AUTODEFINITS

Hi ha molts algorismes d'aprenentatge automàtic possibles per aplicar a una pregunta concreta. El problema és quin algorisme o hiperparàmetres són els millors per a cada plantejament. Una nova metodologia anomenada AutoML consisteix en tècniques que poden seleccionar automàticament el model adequat i els seus hiperparàmetres associats per optimitzar el rendiment i la fiabilitat de les prediccions resultants [111]. Que l'algoritme s'autodefineixi, pot proporcionar un model humà-agnòstic que no sigui propens als biaixos i supòsits lligats a cada decisió que ha de prendre l'expert a l'hora de definir un model. Ja s'ha

provat amb èxit en la identificació de mesures de fenotipat digital més rellevants per als símptomes negatius en trastorns psicòtics [112].

6.3.2. COMBINAR EL CONEIXEMENT D'ALTRES FONTS

En altres dominis, com en la visió per computador, existeixen grans conjunts de dades per a propòsits generals, com ara ImageNet [113]. Però en neuroimatge no és tan fàcil aconseguir un conjunt de dades d'aquesta mida. Aquí és on apareix una tècnica anomenada Aprenentatge per Transferència. Aquest enfocament pot extreure informació obtinguda de grans conjunts de dades de propòsit general i utilitzar aquesta informació per millorar la creació de petits models sobre petites bases de dades [92]. Aquesta tècnica ja s'ha provat per millorar la classificació de la malaltia d'Alzheimer [114]. Així i tot, ara com ara, pràcticament no s'ha fet servir en molts altres dominis com pot ser l'estimació de riscos.

6.3.3. INTEL·LIGÈNCIA ARTIFICIAL INTERPRETABLE

Algorismes nous com l'Aprenentatge profund (Deep Learning) se solen considerar "caixes negres" perquè les decisions de les xarxes no són fàcilment interpretables pels humans. La intel·ligència artificial Explicable (XAI) busca proporcionar una solució fàcilment comprensible. Per exemple, en xarxes neuronals altament complexes utilitzades per a la classificació basada en ressonància magnètica, no és fàcil saber quins vòxels s'han utilitzat per classificar entre grups; XAI proporcionaria un mapa de calor indicant quines eren les zones o vòxels més rellevants utilitzats en la classificació, proporcionant informació sobre el funcionament de la xarxa [115]. Un enfocament és la propagació de la rellevància per capes (LRP), que produeix mapes de calor de la contribució de cada vòxel al resultat final de la classificació per cada subjecte. Quan es va provar en la malaltia d'Alzheimer, els vòxels reportats en el mapa de calor concordaven amb zones que se sap que solen tenir anomalies en aquesta malaltia [82]. També s'ha aplicat a l'esclerosi múltiple, on les lesions es distribueixen per tot el cervell. Els mapes de calor individuals corresponien a les mateixes lesions i a zones de substància grisa i blanca no lesionada com el tàlem, que són marcadors de ressonància

magnètica convencionals [83]. En un estudi on els autors van fer servir mapes de característiques de textura per classificar els participants amb SZ, pacients amb MD i HC, LRP va mostrar quines zones van contribuir a la classificació de l'algorisme d'aprenentatge profund [116]. Un altre enfocament interessant per determinar quines regions contribueixen més a la classificació consisteix a substituir les regions cerebrals per altres sanes generades mitjançant autoencodificadors variacionals i després veure com canvia el rendiment [117].

Disposar d'eines comprensibles per als humans facilitaria que els investigadors, els metges i la població general hi creguessin.

6.3.4. APRENENTATGE FEDERAT

Un obstacle en la compartició de dades per crear conjunts de dades més grans de ressonància magnètica és la preocupació per la privadesa i la confidencialitat. I una altra limitació és que tot i tenir grans bases de dades d'imatges, moltes imatges no estan etiquetades i, per tant, limiten l'aprenentatge del model. Per a etiquetar les imatges, un radiòleg format ha d'inspeccionar les imatges, cosa que pot requerir molt de temps. Tots dos problemes es poden resoldre mitjançant l'aprenentatge federat, ja que permet entrenar algoritmes sense que les dades hagin de sortir de cada centre. Es proporciona un algorisme a tots els centres i s'aplica localment a cada lloc. Un cop l'algoritme extreu la informació, aquest coneixement s'ajunta. Mitjançant aquest enfocament, no es comparteixen dades privades i tots els centres poden ajudar en el procés, fins i tot si la seva base de dades etiquetada és petita. L'aprenentatge federat és una tècnica prometedora que defensa la privacitat dels pacients i facilita la cooperació entre centres sanitaris [118].

6.3.5. UN ENFOCAMENT MULTIMODAL

Se sap que l'esquizofrènia i altres trastorns mentals són causats per una combinació de factors genètics, anatòmics i ambientals. Per tant, les prediccions de resultats futurs o la detecció precoç de subjectes en risc poden beneficiar-se d'enfocaments multimodals, per exemple, combinant factors genètics i neuroanatòmics. De fet, molts estudis ja utilitzen un enfocament

multimodal [119]. No obstant això, el principal problema és que encara no està clar quina combinació de factors prediuen millor el resultat i com combinar-los.

7. DISCUSSIÓ

Aquesta tesi està centrada en un article principal (Article 2) [96], que tenia per objectiu desenvolupar un mètode per poder combinar dades clíniques i dades de RM cerebral per poder obtenir una estimació del risc que pot tenir un pacient a recaure d'acord amb la informació basal. Com que l'obtenció de dades clíniques suposa un esforç tan de temps com econòmic molt grans, s'ha de recórrer a col·laboracions multicèntriques, i és per això que per a aquest estudi les dades provenen de diferents centres. Les dades multicèntriques tenen un biaix potencial associat al centre, i és per això que l'altre article principal (Article 1) [92] juntament amb l'article adjunt [92] pretenen reduir al màxim aquest biaix. El primer article de la tesi (Article 1) [96] se centra a reduir el biaix residual que pot quedar en calcular el rendiment dels models, mentre que l'article adjunt [91] pretén gestionar l'efecte multicèntric harmonitzant les dades en el moment previ a la creació dels models.

Article 1: Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site

En aquest primer estudi [92], que complementa al segon [96], hem mostrat dues formes de controlar l'efecte del centre en estudis multicèntrics. En aquest cas, però, no en l'etapa de la creació del model, sinó en el moment d'estimar com de bé funciona. Hem inclòs dos mètodes per a fer-ho, el primer que consisteix a tractar les dades de cada centre com si fossin estudis diferents i llavors fer una metaanàlisi o el segon, que és incloure la variable del centre com una covariable. En l'estudi, també proporcionem diferents indicacions en funció de la mesura de rendiment utilitzada, i també facilitem un paquet d'R per a aplicar-ho de forma simple.

Per mostrar-ne la importància i també com es pot utilitzar, hem ofert simulacions que altres investigadors podrien utilitzar per comprovar-ne el funcionament, i també un exemple amb dades de RM reals d'una base de dades disponible a internet.

En l'estudi, vam demostrar com utilitzant qualsevol dels dos mètodes, els efectes del centre eren menors i força més petits que els biaixos observats en el cas on no es controlaven. En els exemples en dades simulades, vam poder mostrar que treure aquest efecte pot tenir comportaments diferents en funció de com siguin les dades. En alguns casos podria ser que l'efecte del centre condicioni la pregunta que pretenem respondre amb el model, i, per tant,

podria ser que estiméssim un rendiment superior al real, pel simple fet que les principals diferències vinguin de l'origen d'aquestes dades. També pot passar el cas contrari, que un model tingui un rendiment correcte, però el fet de no tenir en compte el centre, faci que el rendiment estimat acabi essent inferior al que té realment. En l'article hem explicat pas a pas i de forma pretesament pràctica com es poden produir aquests efectes i les diferents formes de tractar-los.

Com a limitació, podríem considerar que centrar l'efecte a només el centre, pot limitar la representativitat del model a la població general, ja que només tindrà en compte les diferències calculades en aquests centres. En alguns casos podria ser necessari controlar a més per altres efectes. Una altra limitació és que seria possible que els efectes del centre i els efectes reals a les dades siguin col·lineals, cosa que faria que la separació d'aquests efectes seria impossible amb els nostres mètodes mostrats.

Per acabar, voldria remarcar la importància que pot tenir el fet de controlar els efectes del centre en l'etapa d'estimació del rendiment del model. I si bé en l'àmbit teòric és impossible extreure tots els efectes potencials que pot tenir el centre, sí que és altament recomanable extreure el màxim possible d'efecte per tal de poder proporcionar estimacions més realistes. Precisament per això, per fer el pas més senzill, hem creat un paquet de molt fàcil ús i que permet fer aquest ajustament de forma molt ràpida, el paquet anomenat "multisite.accuracy".

Article 2: Combining MRI and clinical data to detect high relapse risk after the first episode of psychosis

En aquest estudi, vam crear una eina d'aprenentatge automàtic basada en imatges de RM cerebral per tal de detectar aquells pacients amb un risc alt de recaiguda després de patir un primer episodi psicòtic. Per a fer-ho vam utilitzar una cohort de 227 subjectes que havien patit un PEP. El model va ser capaç de detectar correctament els pacients en alt risc de recaiguda. El perill de recaiguda va ser 4,58 vegades superior en subjectes estimats d'estar en alt risc en comparació als que havien estat detectats com a baix risc. És per això que aquest estudi va aconseguir l'objectiu de crear una eina que pugui ajudar al professional aportant-li una

informació extra que pot ser valuosa. Aquesta informació és especialment útil en les etapes primerenques de la malaltia, ja que permet al professional clínic tenir una eina que li indiqui una informació que li permetria ajustar el tractament profilàctic, que d'altra manera només es podria ajustar després de diverses recaigudes del pacient. Si la prevenció prèvia a l'ajustament correcte del tractament és dèbil, això podria implicar que els pacients en un risc alt de recaiguda podrien experimentar diverses recaigudes; per contra, els pacients que estiguin classificats com a baix risc podrien experimentar efectes adversos a la medicació si s'utilitzés un tractament preventiu massa fort. De totes maneres, qualsevol ajustament del tractament sempre hauria de seguir el principi de "primer no perjudicar", especialment tenint en compte que en la nostra cohort el 85% dels pacients estimats com a alt risc, no van recaure.

Tot i que els resultats són esperançadors, vull fer un incís, ja que aquesta eina desenvolupada per a la tesi, caldria ser sotmesa a una validació extensa prèviament a qualsevol recomanació per ser utilitzada en la pràctica clínica. Desafortunadament, és comú que els resultats d'un estudi no es puguin replicar quan es pretén aplicar a unes noves dades [120], i el nostre estudi podria patir el mateix problema. Per a poder facilitar aquesta replicació i seguint les recomanacions d'estudis previs [121], hem decidit proporcionar una eina en línia senzilla per a facilitar a altres investigadors comprovar els resultats en les seves pròpies dades. Dit això, crec que aquest estudi pot proporcionar un nou pas cap a una potencial aplicació de les tècniques basades en l'aprenentatge automàtic en la pràctica clínica, sempre entesa com a una eina de suport a l'especialista.

També és important, tenir en compte que podria ser útil complementar aquesta eina que proporcionem amb altres tècniques que permetin actualitzar el risc de recaiguda durant el seguiment, ja que per exemple en un estudi previ ja s'ha demostrat que en altres trastorns, com ara el trastorn bipolar, els subjectes en què en un any no han tingut cap recaiguda, el seu risc de recaure a un any vista és molt inferior a altres subjectes [122].

Un fet diferencial d'aquest estudi que crec que cal tenir en compte, és que ens hem centrat a detectar els pacients amb un primer episodi psicòtic que es troben en un alt risc de recaiguda. La majoria d'estudis fins ara s'havien centrat a trobar controls sans amb risc de tenir algun episodi psicòtic [88,123]. I més concretament, d'estudis centrats a utilitzar dades de RM cerebral

per predir el risc de recaiguda en pacients amb un primer episodi psicòtic no se n'han realitzat que coneguem.

En la població general, tal com he mencionat, sí que s'han realitzat molts estudis per tal de detectar aquells subjectes en risc incrementat de patir psicosi, i algunes eines requereixen introduir poca informació per tal de fer una avaluació inicial. N'és un exemple l'eina de detecció Psychosis Polirisk Score (PSS) [124] que només demanant pocs factors de risc ja ha demostrat en un estudi pilot de proporcionar un funcionament adequat per al món real [125].

En aquest estudi, vàrem seguir un pas previ per a escollir els paràmetres de l'algorisme d'aprenentatge automàtic. Per a fer-ho, vam optar per agafar dues bases de dades completament separades de les dades on aplicaríem el model final. D'aquesta manera aconseguíem que la decisió dels paràmetres quedés completament al marge de les dades usades finalment. Aquests paràmetres consistien en incorporar dades de substància grisa i blanca, conjuntament amb les seves imatges modulades, amb un suavitzat amb un nucli σ=4 mm, i fer servir un procés d'"ensemble learning" a nivell de vòxel. A més, vam considerar augmentar la mida del vòxel per reduir la complexitat dels càlculs i així reduir el temps de processat, i vam poder constatar que aquesta reducció no afectava especialment el rendiment dels models, però sí que contribuïa a una millora important en el temps de processament dels models. Aquest pas, també crec que és rellevant remarcar-lo, ja que els paràmetres trobats són extrapolables a altres estudis centrats en la predicció on s'utilitzin tècniques d'aprenentatge automàtic sobre dades de RM.

Tot i l'alegria que ens desperten els resultats trobats, aquest estudi no està exempt de tenir limitacions importants. Primer de tot, tot i que és comú de qualsevol estudi on s'hagi de fer un mostreig de pacients, cal remarcar que la mostra obtinguda pot variar de la mostra de pacients al món real, ja que no s'han inclòs aquells pacients que no han volgut participar en l'estudi o que no complien els requisits d'inclusió. Segon, per molt que tinguem una mostra de 227 pacients seguits durant entre 18 i 24 mesos, només hi ha hagut 16 recaigudes. Això és inferior a estudis anteriors [122,126,127]. Hem pogut comprovar que la diferència en el percentatge de recaigudes és degut al fet que el criteri que hem seguit per definir una recaiguda és més estricte que els estudis anteriors, on consideraven com a recaiguda una hospitalització. Hem intentat replicar aquest criteri de recaiguda, però no ens ha estat

possible aconseguir la informació de les hospitalitzacions per a tots els centres de la nostra mostra. Tercer, la significació estadística és limitada, i això ve probablement donat pel nombre limitat de recaigudes en la mostra. Hem calculat, malgrat això, que el poder estadístic per detectar un risc de recaiguda de 4,5 tal com hem assolit nosaltres seria d'un 72%, força propera al 80% que se sol requerir en el moment de calcular la mida de mostra. Quart, algorismes més avançats com els basats en xarxes neuronals podrien detectar patrons més complexos en les dades, però aquests solen necessitar mostres més grans per tal de tenir un rendiment acceptable. Cinquè, el llindar que hem considerat per classificar pacients en alt o baix risc ha estat tenir HR superior o inferior a 1 respectivament. Aquest llindar l'hem escollit per simplicitat, però podria variar en futurs estudis de validació, i situar-se en un altre valor més bo. Per últim, no hem pogut incloure alguna informació que podria ser rellevant per a la predicció, com és l'adherència a la medicació, el funcionament premòrbid o el temps que passa entre els primers símptomes psicòtics i l'inici del tractament (DUP).

Finalment, aquest estudi representa un pas més cap a l'aplicació translacional dels mètodes de neuroimatge en salut mental. Contràriament a la tendència majoritària de fer servir els models de predicció en neuroimatge per classificar entre diagnòstics, nosaltres ens hem centrat més en la utilitat clínica que podria tenir un mètode per estimar l'evolució individual d'un pacient, més concretament el risc que torni a patir un brot psicòtic. Aquesta eina, en lloc de trobar biomarcadors robustos, se centra principalment a ser útil, és a dir en proporcionar una eina al professional clínic que li pugui ser útil per als seus pacients. I tal com diu el títol de la tesi, els resultats d'aquest treball poden permetre fer un pas més cap a la medicina personalitzada en els trastorns mentals.

8. CONCLUSIONS

1. Quan es crea un model de predicció amb dades que provenen de diferents centres, existeixen efectes del centre que esbiaixen l'estimació de la precisió del model, encara que s'hagin usat mètodes per eliminar les diferències entre centres.

2. Existeixen com a mínim dos mètodes per evitar aquest biaix: un basat a estimar la precisió separadament per a cada centre i després meta-analitzar, i un altre basat a incloure el centre com a covariable en l'estimació de la precisió.

3. Proporcionem a la comunitat científica un paquet d'R amb aquests dos mètodes per tal que altres grups puguin estimar la precisió dels seus models de predicció amb dades multicèntriques sense el biaix relacionat amb els efectes del centre.

4. Es pot aconseguir un augment de precisió dels models de predicció basats en dades de ressonància magnètica cerebral estructural mitjançant la combinació de substància blanca i grisa modulada i no modulada, així com mitjançant l'ús de mètodes de mitjana de meitats del cervell (voxel-based ensemble).

5. En els models de predicció basats en dades de ressonància magnètica cerebral estructural, el submostreig de vòxels disminueix substancialment el cost computacional sense afectar la precisió.

6. Proporcionem a la comunitat científica un paquet d'R per a crear fàcilment models de predicció a partir de dades clíniques i de ressonància magnètica cerebral estructural amb les optimitzacions descrites.

7. En aquesta tesi s'ha creat un model de predicció que permet estimar el risc de recaiguda després d'un primer episodi psicòtic a partir de dades clíniques i de ressonància magnètica cerebral estructural.

8. La precisió de les prediccions és major en aquest model, que combina dades de ressonància magnètica cerebral estructural i dades clíniques augmenta la precisió de les prediccions, que en els models basats només en dades de ressonància magnètica o basats només en dades clíniques.

9. Per tant, l'anatomia cerebral i la clínica aporten informació sobre l'evolució futura d'una persona que ha sofert un primer episodi psicòtic.

10. Proporcionem a la comunitat científica les fórmules i una web per facilitar la replicació del model de predicció creat en aquesta tesi per altres grups.

11. En cas que el model de predicció fos exitosament replicat, podria esdevenir una eina útil en la pràctica clínica en aportar de forma precoç informació que podria ajudar a individualitzar i optimitzar el tractament.

12. Aquesta tesi pretén, per tant, ser un pas més cap a la medicina personalitzada en salut mental i la psiquiatria de precisió.

BIBLIOGRAFIA

1. Isasi, A. G., Echeburúa, E., Limiñana, J. M. & González-Pinto, A. How effective is a psychological intervention program for patients with refractory bipolar disorder? A randomized controlled trial. *J. Affect. Disord.* **126**, 80–87 (2010).
2. Torrent, C. *et al.* Efficacy of functional remediation in bipolar disorder: A multicenter randomized controlled study. *Am. J. Psychiatry* **170**, 852–859 (2013).
3. Keck, P. E. Monitoring pharmacotherapy response, safety, and tolerability to enhance adherence in bipolar disorder. *The Journal of clinical psychiatry* vol. 75 Preprint at https://doi.org/10.4088/JCP.13010tx4c (2014).
4. Popovic, D. *et al.* Polarity index of psychological interventions in maintenance treatment of bipolar disorder. *Psychotherapy and Psychosomatics* vol. 82 292–298 Preprint at https://doi.org/10.1159/000348447 (2013).
5. Brown, E., Bedi, G., McGorry, P. & O'Donoghue, B. Rates and Predictors of Relapse in First-Episode Psychosis: An Australian Cohort Study. *Schizophr. Bull. Open* **1**, (2020).
6. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).
7. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).
8. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
9. Tin Kam Ho. Random decision forests. *Proc. 3rd Int. Conf. Doc. Anal. Recognit.* **1**, 278–282 (1995).

10. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat. 2015 5217553* **521**, 436–444 (2015).

11. Koutsouleris, N. *et al.* Disease Prediction in the At-Risk Mental State for Psychosis Using Neuroanatomical Biomarkers: Results From the FePsy Study. *Schizophr. Bull.* **38**, 1234–1246 (2012).

12. Mikolas, P. *et al.* Machine learning classification of first-episode schizophrenia spectrum disorders and controls using whole brain white matter fractional anisotropy. *BMC Psychiatry* **18**, (2018).

13. Oh, J., Oh, B. L., Lee, K. U., Chae, J. H. & Yun, K. Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. *Front. Psychiatry* **11**, 1–11 (2020).

14. Liu, S. *et al.* Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans. Biomed. Eng.* **62**, 1132–1140 (2015).

15. Schnack, H. G. *et al.* Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *NeuroImage* **84**, 299–306 (2014).

16. Talpalaru, A., Bhagwat, N., Devenyi, G. A., Lepage, M. & Chakravarty, M. M. Identifying schizophrenia subgroups using clustering and supervised learning. *Schizophr. Res.* **214**, 51–59 (2019).

17. Lai, J. W., Ang, C. K. E., Rajendra Acharya, U. & Cheong, K. H. Schizophrenia: A survey of artificial intelligence techniques applied to detection and classification. *Int. J. Environ. Res. Public. Health* **18**, 1–20 (2021).

18. Andreou, C. & Borgwardt, S. Structural and functional imaging markers for susceptibility to psychosis. *Mol. Psychiatry* **25**, 2773–2785 (2020).

19. Koutsouleris, N. *et al.* Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry* **75**, 1156–1172 (2018).

20. Rashid, B. & Calhoun, V. Towards a brain-based predictome of mental illness. *Hum. Brain Mapp.* **41**, 3468–3535 (2020).

21. Arrasate, M. *et al.* Prognostic Value of Affective Symptoms in First-Admission Psychotic Patients. *Int. J. Mol. Sci.* **17**, (2016).

22. Hui, C. L. M. *et al.* Predicting first-episode psychosis patients who will never relapse over 10 years. *Psychol. Med.* **49**, 2206–2214 (2019).

23. Wunderink, L. *et al.* Negative symptoms predict high relapse rates and both predict less favorable functional outcome in first episode psychosis, independent of treatment strategy. *Schizophr. Res.* **216**, 192–199 (2020).

24. Bhattacharyya, S. *et al.* Individualized prediction of 2-year risk of relapse as indexed by psychiatric hospitalization following psychosis onset: Model development in two first episode samples. *Schizophr. Res.* **228**, 483–492 (2021).

25. Bowtell, M. *et al.* Rates and predictors of relapse following discontinuation of antipsychotic medication after a first episode of psychosis. *Schizophr. Res.* **195**, 231–236 (2018).

26. Bergé, D. *et al.* Predictors of Relapse and Functioning in First-Episode Psychosis: A Two-Year Follow-Up Study. *Psychiatr. Serv.* **67**, 227–233 (2016).

27. Schoeler, T. *et al.* Poor medication adherence and risk of relapse associated with continued cannabis use in patients with first-episode psychosis: a prospective analysis. *Lancet Psychiatry* **4**, 627–633 (2017).

28. Kopczynska, M. *et al.* Complement system biomarkers in first episode psychosis. *Schizophr. Res.* **204**, 16–22 (2019).

29. Laskaris, L. *et al.* Investigation of peripheral complement factors across stages of psychosis. *Schizophr. Res.* **204**, 30–37 (2019).

30. Harrisberger, F. *et al.* Impact of polygenic schizophrenia-related risk and hippocampal volumes on the onset of psychosis. *Transl. Psychiatry* **6**, e868 (2016).

31. Ranlund, S. *et al.* A polygenic risk score analysis of psychosis endophenotypes across brain functional, structural, and cognitive domains. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **177**, 21–34 (2018).

32. Vassos, E. *et al.* An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biol. Psychiatry* **81**, 470–477 (2017).

33. Schmidt, A. *et al.* Improving Prognostic Accuracy in Subjects at Clinical High Risk for Psychosis: Systematic Review of Predictive Models and Meta-analytical Sequential Testing Simulation. *Schizophr. Bull.* **43**, 375–388 (2017).

34. Cahn, W. *et al.* Brain volume changes in the first year of illness and 5-year outcome of schizophrenia. *Br. J. Psychiatry* **189**, 381–382 (2006).

35. Dazzan, P. *et al.* 31.3 CLINICAL UTILITY OF MRI SCANNING IN FIRST EPISODE PSYCHOSIS. *Schizophr. Bull.* **44**, S50–S51 (2018).

36. Nieuwenhuis, M. *et al.* Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. *NeuroImage* **145**, 246–253 (2017).

37. Dazzan, P. *et al.* Magnetic resonance imaging and the prediction of outcome in first-episode schizophrenia: a review of current evidence and directions for future research. *Schizophr. Bull.* **41**, 574–83 (2015).

38. Korda, A. I., Andreou, C. & Borgwardt, S. Pattern classification as decision support tool in antipsychotic treatment algorithms. *Exp. Neurol.* **339**, 113635 (2021).

39. Correll, C. U. & Schooler, N. R. Negative Symptoms in Schizophrenia: A Review and Clinical Guide for Recognition, Assessment, and Treatment. *Neuropsychiatr. Dis. Treat.* **16**, 519–534 (2020).

40. Rodrigues, R. & Anderson, K. K. The traumatic experience of first-episode psychosis: A systematic review and meta-analysis. *Schizophr. Res.* **189**, 27–36 (2017).

41. Addington, J., Coldham, E. L., Jones, B., Ko, T. & Addington, D. The first episode of psychosis: the experience of relatives. *Acta Psychiatr. Scand.* **108**, 285–289 (2003).

42. Jones, P. B. Adult mental health disorders and their age at onset. *Br. J. Psychiatry* **202**, s5–s10 (2013).

43. Solmi, M. *et al.* Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol. Psychiatry* **17**, 22 (2021).

44. Suvisaari, J. *et al.* Is It Possible to Predict the Future in First-Episode Psychosis? *Frontiers in Psychiatry* vol. 9 580 Preprint at https://doi.org/10.3389/fpsyt.2018.00580 (2018).

45. Salvatore, P. *et al.* McLean-Harvard International First-Episode Project. *J. Clin. Psychiatry* **70**, 458–466 (2009).

46. Alvarez-Jimenez, M. *et al.* Risk factors for relapse following treatment for first episode psychosis: A systematic review and meta-analysis of longitudinal studies. *Schizophr. Res.* **139**, 116–128 (2012).

47. Birchwood, M., Todd, P. & Jackson, C. Early intervention in psychosis: The critical period hypothesis. *Br. J. Psychiatry* **172**, 53–59 (1998).

48. Fusar-Poli, P., McGorry, P. D. & Kane, J. M. Improving outcomes of first-episode psychosis: an overview. *World Psychiatry* **16**, 251–265 (2017).

49. Howes, O. D. *et al.* The clinical significance of duration of untreated psychosis: an umbrella review and random-effects meta-analysis. *World Psychiatry* **20**, 75–95 (2021).

50. Kay, S. R., Fiszbein, A. & Opler, L. A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* **13**, 261–276 (1987).

51. Andreasen, N. C. *et al.* Remission in schizophrenia: proposed criteria and rationale for consensus. *Am. J. Psychiatry* **162**, 441–449 (2005).

52. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).

53. Yassin, W. *et al.* Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. *Transl. Psychiatry* **10**, 1–11 (2020).

54. Salvador, R. *et al.* Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLOS ONE* **12**, e0175683 (2017).

55. Zarogianni, E., Moorhead, T. W. J. & Lawrie, S. M. Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. *NeuroImage Clin.* **3**, 279 (2013).

56. Yassin, W. *et al.* Gray matter alterations in schizophrenia high-risk youth and early-onset schizophrenia: a review of structural MRI findings. *undefined* **10**, (2013).

57. Zarogianni, E., Storkey, A. J., Johnstone, E. C., Owens, D. G. C. & Lawrie, S. M. Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophr. Res.* **181**, 6–12 (2017).

58. Zhu, F. *et al.* Functional asymmetry of thalamocortical networks in subjects at ultra-high risk for psychosis and first-episode schizophrenia. *Eur. Neuropsychopharmacol.* **29**, 519–528 (2019).

59. Lalousis, P. A. *et al.* Heterogeneity and Classification of Recent Onset Psychosis and Depression: A Multimodal Machine Learning Approach. *Schizophr. Bull.* **47**, 1130–1140 (2021).

60. Gould, I. C. *et al.* Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach. *NeuroImage Clin.* **6**, 229–236 (2014).

61. Wenzel, J. *et al.* Cognitive subtypes in recent onset psychosis: distinct neurobiological fingerprints? *Neuropsychopharmacol. 2021 468* **46**, 1475–1483 (2021).

62. Guo, Y., Qiu, J. & Lu, W. Support Vector Machine-Based Schizophrenia Classification Using Morphological Information from Amygdaloid and Hippocampal Subregions. *Brain Sci.* **10**, E562 (2020).

63. Anderson, A. & Cohen, M. S. Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. *Front. Hum. Neurosci.* **7**, (2013).

64. Wang, J., Ke, P., Zang, J., Wu, F. & Wu, K. Discriminative Analysis of Schizophrenia Patients Using Topological Properties of Structural and Functional Brain Networks: A Multimodal Magnetic Resonance Imaging Study. *Front. Neurosci.* **15**, 785595 (2021).

65. Steardo, L. *et al.* Application of support vector machine on fmri data as biomarkers in schizophrenia diagnosis: A systematic review. *Front. Psychiatry* **11**, 1–9 (2020).

66. Du, Y., Fu, Z. & Calhoun, V. D. Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. *Front. Neurosci.* **12**, (2018).

67. Mika, S., Rätsch, G., Weston, J., Schölkopf, B. & Müller, K.-R. Fisher Discriminant Analysis With Kernels. (1999).

68. Karageorgiou, E. *et al.* Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. *Neuroinformatics* **9**, 321–333 (2011).

69. Kasparek, T. *et al.* Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. *Psychiatry Res. Neuroimaging* **191**, 174–181 (2011).

70. Takayanagi, Y. *et al.* Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **34**, 10–17 (2010).

71. Zanetti, M. V. *et al.* Neuroanatomical pattern classification in a population-based sample of first-episode schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **43**, 116–125 (2013).

72. Kawasaki, Y. *et al.* Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *NeuroImage* **34**, 235–242 (2007).

73. Nakamura, K. *et al.* Multiple Structural Brain Measures Obtained by Three-Dimensional Magnetic Resonance Imaging To Distinguish Between Schizophrenia Patients and Normal Subjects. *Schizophr. Bull.* **30**, 393–404 (2004).

74. Ota, M. *et al.* Discrimination of female schizophrenia patients from healthy women using multiple structural brain measures obtained with voxel-based morphometry. *Psychiatry Clin. Neurosci.* **66**, 611–617 (2012).

75. Santos, P. E. *et al.* Exploring the knowledge contained in neuroimages: Statistical discriminant analysis and automatic segmentation of the most significant changes. *Artif. Intell. Med.* **49**, 105–115 (2010).

76. Winterburn, J. L. *et al.* Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophr. Res.* **214**, 3–10 (2019).

77. Greenstein, D., Malley, J. D., Weisinger, B., Clasen, L. & Gogtay, N. Using Multivariate Machine Learning Methods and Structural MRI to Classify Childhood Onset Schizophrenia and Healthy Controls. *Front. Psychiatry* **3**, 53 (2012).

78. Schwarz, E. *et al.* Reproducible grey matter patterns index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder. *Transl. Psychiatry* **9**, 1–13 (2019).

79. Shahab, S. *et al.* Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* **44**, 898–906 (2019).

80. Ullah, F. *et al.* Brain MR Image Enhancement for Tumor Segmentation Using 3D U-Net. *Sensors* **21**, 7528 (2021).

81. Shoeibi, A. *et al.* Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Comput. Biol. Med.* **136**, 104697 (2021).

82. Böhle, M., Eitel, F., Weygandt, M. & Ritter, K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front. Aging Neurosci.* **10**, 194 (2019).

83. Eitel, F. *et al.* Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage Clin.* **24**, (2019).

84. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–220 (1972).

85. Zaks, N. *et al.* Sleep Disturbance in Individuals at Clinical High Risk for Psychosis. *Schizophr. Bull.* **48**, 111–121 (2022).

86. Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Front. Psychiatry* **7**, (2016).

87. Radua, J. & Carvalho, A. F. Route map for machine learning in psychiatry: Absence of bias, reproducibility, and utility. *Eur. Neuropsychopharmacol.* **50**, 115–117 (2021).

88. Rosen, M. *et al.* Towards clinical application of prediction models for transition to psychosis: A systematic review and external validation study in the PRONIA sample. *Neurosci. Biobehav. Rev.* **125**, 478–492 (2021).

89. Navarro, C. L. A. *et al.* Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* **375**, n2281 (2021).

90. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci. 2009 125* **12**, 535–540 (2009).

91. Radua, J. *et al.* Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage* **218**, (2020).

92. Solanes, A. *et al.* Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. *Psychiatry Res. Neuroimaging* **314**, 111313 (2021).

93. Kim, Y. K. & Na, K. S. Application of machine learning classification for structural brain MRI in mood disorders: Critical review from a clinical perspective. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **80**, 71–80 (2018).

94. Mechelli, A. & Vieira, S. From models to tools: clinical translation of machine learning studies in psychosis. *Npj Schizophr.* **6**, 4 (2020).

95. Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* **29**, 102–127 (2019).

96. Solanes, A. *et al.* Combining MRI and clinical data to detect high relapse risk after the first episode of psychosis. *Schizophr. Heidelb. Ger.* **8**, 100 (2022).

97. Chen, B. & Benedetti, A. Quantifying heterogeneity in individual participant data meta-analysis with binary outcomes. *Syst. Rev.* **6**, 243 (2017).

98. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

99. Tustison, N. J. *et al.* Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage* **99**, 166–179 (2014).

100. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).

101. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).

102. Van Erp, T. G. M. *et al.* Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* **21**, 547–553 (2016).

103. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *NeuroImage* **92**, 381–397 (2014).

104. Nakagawa, S. *et al.* Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods Ecol. Evol.* **6**, 143–152 (2015).

105. Marcus, D. S. *et al.* Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**, 1498–1507 (2007).

106. DeLisi, L. E. *et al.* Cerebral ventricular enlargement as a possible genetic marker for schizophrenia. *Psychopharmacol. Bull.* **21**, 365–367 (1985).

107. Greenshtein, E. & Ritov, Y. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988 (2004).

108. Gallardo-Ruiz, R., Crespo-Facorro, B., Setién-Suero, E. & Tordesillas-Gutierrez, D. Long-Term Grey Matter Changes in First Episode Psychosis: A Systematic Review. *Psychiatry Investig.* **16**, 336 (2019).

109. Li, R.-R. *et al.* Altered functional connectivity strength and its correlations with cognitive function in subjects with ultra-high risk for psychosis at rest. *CNS Neurosci. Ther.* **24**, 1140–1148 (2018).

110. Abrol, A. *et al.* Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* **12**, 1–17 (2021).

111. Dafflon, J. *et al.* An automated machine learning approach to predict brain age from cortical anatomical measures. *Hum. Brain Mapp.* **41**, 3555–3566 (2020).

112. Narkhede, S. M. *et al.* Machine Learning Identifies Digital Phenotyping Measures Most Relevant to Negative Symptoms in Psychotic Disorders: Implications for Clinical Trials. *Schizophr. Bull.* sbab134 (2021) doi:10.1093/schbul/sbab134.

113. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009). doi:10.1109/CVPR.2009.5206848.

114. Hon, M. & Khan, N. M. Towards Alzheimer's disease classification through transfer learning. *Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017* **2017-Janua**, 1166–1169 (2017).

115. Fellous, J. M., Sapiro, G., Rossi, A., Mayberg, H. & Ferrante, M. Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation. *Front. Neurosci.* **13**, 1–14 (2019).

116. Korda, A. I. *et al.* Identification of voxel-based texture abnormalities as new biomarkers for schizophrenia and major depressive patients using layer-wise relevance

propagation on deep learning decisions. *Psychiatry Res. Neuroimaging* **313**, 111303 (2021).

117. Uzunova, H., Ehrhardt, J., Kepp, T. & Handels, H. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. in *Medical Imaging 2019: Image Processing* (eds. Angelini, E. D. & Landman, B. A.) 36 (SPIE, 2019). doi:10.1117/12.2511964.

118. Ng, D., Lan, X., Yao, M. M.-S., Chan, W. P. & Feng, M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant. Imaging Med. Surg.* **11**, 852–857 (2021).

119. Ulaş, A. *et al.* Multimodal schizophrenia detection by multiclassification analysis. *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.* **7042 LNCS**, 491–498 (2011).

120. Radua, J. What is the actual accuracy of clinical prediction models? The case of transition to psychosis. *Neurosci. Biobehav. Rev.* **127**, 502–503 (2021).

121. Young, J., Kempton, M. J. & McGuire, P. Using machine learning to predict outcomes in psychosis. *Lancet Psychiatry* **3**, 908–909 (2016).

122. Radua, J., Grunze, H. & Amann, B. L. Meta-Analysis of the Risk of Subsequent Mood Episodes in Bipolar Disorder. *Psychother. Psychosom.* **86**, 90–98 (2017).

123. Smieskova, R. *et al.* Neuroimaging predictors of transition to psychosis–a systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **34**, 1207–1222 (2010).

124. Oliver, D., Radua, J., Reichenberg, A., Uher, R. & Fusar-Poli, P. Psychosis Polyrisk Score (PPS) for the Detection of Individuals At-Risk and the Prediction of Their Outcomes. *Front. Psychiatry* **10**, 174 (2019).

125. Oliver, D. *et al.* Real-world digital implementation of the Psychosis Polyrisk Score (PPS): A pilot feasibility study. *Schizophr. Res.* **226**, 176–183 (2020).

126. Tiihonen, J. *et al.* A nationwide cohort study of oral and depot antipsychotics after first hospitalization for schizophrenia. *Am. J. Psychiatry* **168**, 603–609 (2011).

127. Robinson, D. *et al.* Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. *Arch. Gen. Psychiatry* **56**, 241–247 (1999).

MATERIAL ADJUNT

ARTICLE ADJUNT: INCREASED POWER BY HARMONIZING STRUCTURAL MRI SITE DIFERRENCES WITH THE COMBAT BATCH ADJUSTMENT METHOD IN ENIGMA

NeuroImage 218 (2020) 116956

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA

Joaquim Radua [a,b,c,d,*], Eduard Vieta [b,e,f,g], Russell Shinohara [h,i], Peter Kochunov [j], Yann Quidé [k,l], Melissa J. Green [k,l], Cynthia S. Weickert [k,l,m], Thomas Weickert [k,l], Jason Bruggemann [k,l], Tilo Kircher [n], Igor Nenadić [n], Murray J. Cairns [o,p], Marc Seal [q,r], Ulrich Schall [o,p], Frans Henskens [s], Janice M. Fullerton [be,l], Bryan Mowry [t,u], Christos Pantelis [v,w], Rhoshel Lenroot [k,l,x], Vanessa Cropley [v], Carmel Loughland [o], Rodney Scott [o], Daniel Wolf [y], Theodore D. Satterthwaite [y], Yunlong Tan [z], Kang Sim [aa,ab,ac], Fabrizio Piras [ad], Gianfranco Spalletta [ad,ae], Nerisa Banaj [ad], Edith Pomarol-Clotet [b,af], Aleix Solanes [a,b,af,ag], Anton Albajes-Eizagirre [a,b,af], Erick J. Canales-Rodríguez [b,af,ah,ai], Salvador Sarro [b,af,aj], Annabella Di Giorgio [ak,al], Alessandro Bertolino [al], Michael Stäblein [am], Viola Oertel [am], Christian Knöchel [am], Stefan Borgwardt [an,cc], Stefan du Plessis [ao], Je-Yeon Yun [ap,aq], Jun Soo Kwon [ar,as], Udo Dannlowski [at], Tim Hahn [at], Dominik Grotegerd [at], Clara Alloza [b,au,av], Celso Arango [b,au,av,aw], Joost Janssen [b,au,av], Covadonga Díaz-Caneja [b,au,av,aw], Wenhao Jiang [ax], Vince Calhoun [ay], Stefan Ehrlich [az], Kun Yang [ba], Nicola G. Cascella [ba], Yoichiro Takayanagi [bb,bc], Akira Sawa [bc,bd], Alexander Tomyshev [bf], Irina Lebedeva [bf], Vasily Kaleda [bf], Matthias Kirschner [bg,bh], Cyril Hoschl [bi,cd], David Tomecek [bi,bj,bk], Antonin Skoch [bi,bl], Therese van Amelsvoort [bm], Geor Bakker [bm], Anthony James [bn], Adrian Preda [bo], Andrea Weideman [bo], Dan J. Stein [bp], Fleur Howells [bq,br], Anne Uhlmann [bq,bs], Henk Temmingh [bq,bt], Carlos López-Jaramillo [bu,bv], Ana Díaz-Zuluaga [bu], Lydia Fortea [a], Eloy Martinez-Heras [g,bw,bx], Elisabeth Solana [g,bw,bx], Sara Llufriu [g,bw,bx], Neda Jahanshad [by], Paul Thompson [bz], Jessica Turner [ax], Theo van Erp [ca,cb], ENIGMA Consortium collaborators

[a] *Imaging of Mood- and Anxiety-Related Disorders (IMARD) Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain*
[b] *CIBERSAM, Madrid, Spain*
[c] *Early Psychosis: Interventions and Clinical-detection (EPIC) Lab, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK*
[d] *Department of Clinical Neuroscience, Stockholm Health Care Services, Stockholm County Council, Karolinska Institutet, Stockholm, Sweden*
[e] *Bipolar and depressive disorders, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain*
[f] *Barcelona Bipolar Disorders Program, Institute of Neurosciences, Hospital Clinic de Barcelona, Barcelona, Spain*
[g] *University of Barcelona, Barcelona, Spain*
[h] *Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
[i] *Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, USA*
[j] *Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, MD, USA*
[k] *School of Psychiatry, University of New South Wales, Sydney, NSW, Australia*
[l] *Neuroscience Research Australia, Sydney, NSW, Australia*
[m] *Department of Neuroscience & Physiology, Upstate Medical University, Syracuse, Newyork, NY, USA*
[n] *Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Marburg, Germany*
[o] *University of Newcastle, Newcastle, NSW, Australia*
[p] *Hunter Medical Research Institute, Newcastle, NSW, Australia*
[q] *Murdoch Children's Research Institute, Melbourne, VIC, Australia*
[r] *The University of Melbourne, Australia*
[s] *Health Behaviour Research Group, School of Medicine and Public Health, University of Newcastle, Newcastle, NSW, Australia*

* Corresponding author. Imaging of Mood- and Anxiety-Related Disorders (IMARD) group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
E-mail address: radua@clinic.cat (J. Radua).

https://doi.org/10.1016/j.neuroimage.2020.116956

Available online 26 May 2020

[t] Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia
[u] Queensland Centre for Mental Health Research, The University of Queensland, Brisbane, QLD, Australia
[v] Melbourne Neuropsychiatry Centre, Dept. of Psychiatry, University of Melbourne, Melbourne, VIC, Australia
[w] North Western Mental Health, Melbourne Health, Melbourne, VIC, Australia
[x] University of New Mexico, Albuquerque, NM, USA
[y] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA
[z] Psychiatry Research Center, Beijing Huilongguan Hospital, Beijing, China
[aa] West Region and Research Division, Institute of Mental Health, Singapore, Singapore
[ab] Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[ac] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[ad] Laboratory of Neuropsychiatry, Department of Clinical and Behavioral Neurology, IRCCS Santa Lucia Foundation, Rome, Italy
[ae] Division of Neuropsychiatry, Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, USA
[af] FIDMAG Germanes Hospitalàries Research Foundation, Barcelona, Spain
[ag] Department of Psychiatry and Forensic Medicine, School of Medicine, Autonomous University of Barcelona, Barcelona, Spain
[ah] Department of Radiology, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[ai] Signal Processing Lab (LTS5), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[aj] School of Medicine, Universitat Internacional de Catalunya, Barcelona, Spain
[ak] IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy
[al] Department of Basic Medical Science, Neuroscience and Sense Organs, University of Bari 'Aldo Moro', Bari, Italy
[am] Dept. of Psychiatry, Psychosomatic Medicine and Psychotherapy, Goethe University Frankfurt, Frankfurt, Germany
[an] Department of Psychiatry, University of Basel, Basel, Switzerland
[ao] University of Stellenbosch, Cape Town, Western Province, South Africa
[ap] Seoul National University Hospital, Seoul, Republic of Korea
[aq] Yeongeon Student Support Center, Seoul National University College of Medicine, Seoul, Republic of Korea
[ar] Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea
[as] Department of Brain & Cognitive Sciences, College of Natural Sciences, Seoul National University, Seoul, Republic of Korea
[at] Department of Psychiatry, University of Münster, Münster, Germany
[au] Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio Marañón, Madrid, Spain
[av] Instituto de Investigación Sanitaria Gregorio Marañón (IiSGM), Madrid, Spain
[aw] School of Medicine, Universidad Complutense, Madrid, Spain
[ax] Georgia State University, Atlanta, GA, USA
[ay] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, Atlanta, GA, USA
[az] Technische Universität Dresden, Faculty of Medicine, Division of Psychological and Social Medicine, Dresden, Germany
[ba] Departments of Psychiatry, Johns Hopkins School of Medicine, Baltimore, MD, USA
[bb] Department of Neuropsychiatry, University of Toyama Graduate School of Medicine and Pharmaceutical Sciences, Toyama, Japan
[bc] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[bd] Departments of Psychiatry, Neuroscience, and Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD, USA
[be] School of Medical Sciences, University of New South Wales, Sydney, NSW, Australia
[bf] Mental Health Research Center, Moscow, Russia
[bg] Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric Hospital, University of Zurich, Zurich, Switzerland
[bh] Montreal Neurological Institute, McGill University, Montreal, Canada
[bi] National Institute of Mental Health, Klecany, Czech Republic
[bj] Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic
[bk] Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[bl] MR Unit, Department of Diagnostic and Interventional Radiology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic
[bm] Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands
[bn] Department of Psychiatry, University of Oxford, Oxford, UK
[bo] Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA
[bp] SAMRC Unit on Risk & Resilience in Mental Disorders, Dept of Psychiatry and Neuroscience Institute, University of Cape Town, Cape Town, Western Province, South Africa
[bq] Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, Western Cape, South Africa
[br] Neuroscience Institute, University of Cape Town, Cape Town, Western Cape, South Africa
[bs] Department of Child and Adolescent Psychiatry, Technische Universität Dresden, Dresden, Germany
[bt] Valkenburg Hospital, Observatory, Cape Town, Western Cape, South Africa
[bu] Research Group in Psychiatry GIPSI, Department of Psychiatry, Faculty of Medicine, Universidad de Antioquia, Medellín, Antioquia, Colombia
[bv] Mood Disorders Program, Hospital Universitario San Vicente Fundación, Medellin, Colombia
[bw] Center of Neuroimmunology. Laboratory of Advanced Imaging in Neuroimmunological Diseases. Hospital Clinic de Barcelona, Barcelona, Spain
[bx] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
[by] Imaging Genetics Center, Mark & Mary Stevens Neuroimaging & Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[bz] Imaging Genetics Center, Department of Neurology, University of Southern California, Los Angeles, CA, USA
[ca] Clinical Translational Neuroscience Laboratory, Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA
[cb] Center for the Neurobiology of Learning and Memory, University of California Irvine, 309 Qureshey Research Lab, Irvine, CA, 92697, USA
[cc] Department of Psychiatry and Psychotherapy, University Lübeck, Germany
[cd] Department of Psychiatry and Clinical Psychology, Third Faculty of Medicine, Charles University, Prague, Czech Republic

ARTICLE INFO

Keywords:
Brain
Cortical thickness
Gray matter
Mega-analysis
Neuroimaging
Schizophrenia
Volume

ABSTRACT

A common limitation of neuroimaging studies is their small sample sizes. To overcome this hurdle, the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) Consortium combines neuroimaging data from many institutions worldwide. However, this introduces heterogeneity due to different scanning devices and sequences. ENIGMA projects commonly address this heterogeneity with random-effects meta-analysis or mixed-effects mega-analysis. Here we tested whether the batch adjustment method, ComBat, can further reduce site-related heterogeneity and thus increase statistical power. We conducted random-effects meta-analyses, mixed-effects mega-analyses and ComBat mega-analyses to compare cortical thickness, surface area and subcortical volumes between 2897 individuals with a diagnosis of schizophrenia and 3141 healthy controls from 33 sites. Specifically, we compared the imaging data between individuals with schizophrenia and healthy controls, covarying for age and sex. The use of ComBat substantially increased the statistical significance of the findings as compared to random-

effects meta-analyses. The findings were more similar when comparing ComBat with mixed-effects mega-analysis, although ComBat still slightly increased the statistical significance. ComBat also showed increased statistical power when we repeated the analyses with fewer sites. Results were nearly identical when we applied the ComBat harmonization separately for cortical thickness, cortical surface area and subcortical volumes. Therefore, we recommend applying the ComBat function to attenuate potential effects of site in ENIGMA projects and other multi-site structural imaging work. We provide easy-to-use functions in R that work even if imaging data are partially missing in some brain regions, and they can be trained with one data set and then applied to another (a requirement for some analyses such as machine learning).

1. Introduction

After the early reporting of ventricular enlargement in patients with schizophrenia (SCZ) using pneumoencephalography (Huber, 1957), there has been an exponential increase in the number of studies that use imaging techniques to detect brain differences in people with psychiatric disorders. This increase is most evident for studies using magnetic resonance imaging (MRI), probably due to its high resolution and its wide availability around the globe. However, most MRI studies have examined relatively small sample sizes, a limitation that may prevent the detection of true differences (type II errors), and because of the use of liberal thresholds, may even lead to increased detection of false differences (type I errors). Consequently, reports of unreliable, inconsistent and even contradictory results are not uncommon (Radua and Mataix-Cols, 2012).

Collaborative multi-site initiatives provide an opportunity to assemble larger and more diverse groups of subjects, leading to increased power and findings that may be more representative of the general population. Among these initiatives, the ENIGMA (Enhancing Neuro Imaging Genetics through Meta-Analysis; http://enigma.ini.usc.edu) Consortium (Thompson et al., 2014) stands out for including hundreds of groups worldwide and facilitating the sharing of tens of thousands of neuroimages. One great advantage of this consortium is the harmonization of the protocols to pre-process the MRI data, which has decreased the heterogeneity between the sites related to methodological factors. All sites apply the same pre-processing pipelines to obtain thickness and surface area estimates for cortical regions of interest (ROI) and volume estimates for subcortical ROIs; similar harmonized protocols are in use for standardized analysis of diffusion MRI, resting state fMRI and EEG data, as well as various kinds of omics data (GWAS and epigenetic data).

However, even though all sites participating in an ENIGMA project apply the same pre-processing protocol, data from different sites still show relevant methodological heterogeneity due to systematic differences in MRI scanning devices and acquisition sequences. Also, prior studies have reported that the results of the FreeSurfer segmentation process, for morphometric analysis of MRI, can be affected even by using different FreeSurfer versions, workstations or operating systems (Chepkoech et al., 2016; Gronenschild et al., 2012). Most ENIGMA projects address this residual heterogeneity by random-effects meta-analysis (RE-Meta), but estimation and control of heterogeneity in site-aggregated meta-analyses may be suboptimal (Chen and Benedetti, 2017). It is worth noting that a few ENIGMA studies have analyzed shared individual data (rather than site-aggregated statistical data). These "mega-analyses" of individual data considered the "site" as a random factor within a linear mixed-effects model (ME-Mega), and in several cases examined so far, showed higher statistical power than RE-Meta (Boedhoe et al., 2017, 2018; Favre et al., 2019; van Rooij et al., 2018) (Table 1).

Here, we tested whether ME-Mega may be further improved using a recently developed method to control for batch effects. Standard ME-Mega assumes that the error terms follow the same normal distribution at all sites, which is rarely the case as sites usually have different error variances. In addition, both RE-Mega and ME-Meta estimate the heterogeneity of each ROI independently, while it is likely that all ROIs share some heterogeneity. One method that overcomes these issues is ComBat (Johnson et al., 2007), a batch adjustment method developed for genomics data. Fortin and colleagues have shown that ComBat mega-analysis (ComBat-Mega) outperformed other methods for removing the effects of site from cortical thickness data obtained using the ANTs cortical thickness pipeline (Tustison et al., 2014) from a moderately small number of different sites ($\leq$7 sites). Specifically, ComBat decreased scan-related heterogeneity and increased statistical power and reproducibility (Fortin et al., 2018). The current study examines whether this harmonization result can be extended to ENIGMA data obtained using a standardized FreeSurfer pipeline (Dale et al., 1999; Fischl et al., 1999). Moreover, we did not know whether the use of a larger number of sites could minimize the advantages of ComBat-Mega as compared to ME-Mega. To answer these questions, we analyzed the main structural MRI data from the ENIGMA Schizophrenia Working Group using RE-Meta, ME-Mega and ComBat-Mega, and then compared the findings. The RE-Meta of these data have been already published (van Erp et al., 2016, 2018; Wong et al., 2019); in those analyses, individuals with SCZ showed widespread thinner cortex and smaller surface area, as well as smaller hippocampus, amygdala, thalamus and accumbens volumes, and larger pallidum and lateral ventricle volumes.

We hypothesized that ComBat-Mega would show improvements over RE-Meta and ME-Mega in detecting differences between groups of individuals with SCZ and healthy controls (CON), with standard errors of these effects scaling by method: Combat-Mega < ME-Mega < RE-Meta. We further provide the R code (http://enigma.ini.usc.edu/wp-content/uploads/combat_for_ENIGMA_sMRI/combat_for_ENIGMA_sMRI.R) for the application of ComBat harmonization for other ENIGMA mega-analyses or other multi-site structural imaging work even if the imaging data are partially missing in some ROIs (the original ComBat function did not accept missing data).

2. Methods

2.1. Methodological approaches

Before detailing the collection of data and analyses in the present study, we will briefly explain the three methodological approaches. To exemplify the explanation, we will refer to a simple comparison of cortical thickness between groups of individuals with SCZ and CON, after

Table 1
Previous ENIGMA projects that included both mega-analyses and meta-analyses.

	RE-Meta	ME-Mega
Subcortical volumes in obsessive-compulsive disorder (Boedhoe et al., 2017)	↓ in 1 ROI and ↑ in 1 ROI	↓ in 1 ROI and ↑ in 1 ROI
Fractional anisotropy in bipolar disorder (Favre et al., 2019)	↓ in 23 out of 44 ROIs	↓ in 29 out of 44 ROIs
Cortical thickness in obsessive-compulsive disorder (Boedhoe et al., 2018)	No findings	↓ in 2 ROIs
Surface area in obsessive-compulsive disorder (Boedhoe et al., 2018)	↓ in 1 ROI	↓ in 1 ROI
Subcortical volumes in autism spectrum disorder (van Rooij et al., 2018)	↓ in 3 ROIs	↓ in 4 ROIs
Cortical thickness in autism spectrum disorder (van Rooij et al., 2018)	↑ in 3 ROIs and ↓ in 10 ROIs	↑ in 9 ROIs and ↓ in 7 ROIs

Footnote: ROI: region of interest. ME-Mega: mixed-effects mega-analysis; RE-Meta: random-effects meta-analysis.

covarying for effects of age and sex, but the concepts are applicable to other measures and statistical contrasts. We also conducted an alternative analysis covarying for age, sex, and intracranial volume (ICV).

2.1.1. The RE-Meta approach

In the random-effects meta-analysis (RE-Meta), a linear model estimates the difference in cortical thickness between SCZ and CON for each ROI at each site, covarying for age and sex:

$$y_{r,i,j} = \alpha_{r,i} + X_{i,j} \cdot \beta_{r,i} + \varepsilon_{r,i,j}$$

where $y_{r,i,j}$ is the measurement of cortical thickness of the rth ROI from the jth individual of the ith site, $\alpha_{r,i}$ is the estimate overall cortical thickness of the rth ROI from individuals of the ith site, $X_{i,j}$ are the values of the variables (disorder, age, and sex) of the jth individual of the ith site, $\beta_{r,i}$ are the estimates of the coefficients of these variables for the rth ROI from individuals of the ith site, and $\varepsilon_{r,i,j}$ is the error term for the rth ROI in the jth individual of the ith site.

Estimates of coefficients of interest (e.g., $\beta_{r,i,1}$, the difference between SCZ and CON are then pooled to obtain a single estimate for each ROI ($\beta_{r,meta,1}$). A typical method to pool the coefficients is the weighted mean of the coefficient of each site (Radua and Mataix-Cols, 2012):

$$\beta_{r,meta,1} = \sum_{i \in sites} (w_{r,i} \cdot \beta_{r,i,1})$$

where $w_{r,i}$ the weight of ith site for the rth ROI, and is calculated as the inverse of the variance of $\beta_{r,i,1}$, plus the heterogeneity for the rth ROI (τ_r^2):

$$w_{r,i} = \frac{1}{\mathrm{var}(\beta_{r,i,1}) + \tau_r^2}$$

Frequently, the analyst does not use the coefficients but effect sizes, such as Hedges' g (Radua and Mataix-Cols, 2012), but the concept is similar.

Some problems of RE-Meta are that $\beta_{r,i}$ may be poorly estimated in sites with small sample sizes, or that τ_r^2 may be poorly estimated in some scenarios (Chen and Benedetti, 2017).

2.1.2. The ME-Mega approach

In the standard mixed effects mega-analysis (ME-Mega), a linear mixed-effects model is performed on shared individual subject data to estimate the overall difference in cortical thickness between SCZ and CON, for each ROI, covarying for age and sex. This analysis is conducted in a single step, with "site" included in the model as a random factor:

$$y_{r,i,j} = \alpha_r + X_{i,j} \cdot \beta_r + \gamma_{r,i} + \varepsilon_{r,i,j}$$

where α_r is the estimate overall cortical thickness of the rth ROI from all individuals, β_r are the estimates of the coefficients of the variables for the rth ROI from all individuals, and $\gamma_{r,i}$ are the additive effects of the ith site in the rth ROI.

This approach benefits from a more robust estimation of α_r and β_r as it is based on the data from all sites, as well as from a more precise estimation of the heterogeneity. However, it still may have some minor issues. It assumes that the error terms follow the same normal distribution at all sites, which may seldom be the case. We acknowledge that it is possible to create linear mixed-effects models that consider a different variance for each site, but they involve the specification of variance structures for each statistical test, which may substantially complicate the analyses. In addition, the effects of site are estimated independently for each ROI, which may be suboptimal because the effects of site, even if different for each ROI, may still share some commonalities (e.g., an MRI device may yield a better signal contrast than another across the brain).

2.1.3. The ComBat-mega approach

As compared to ME-Mega, the Combat mega-analysis (ComBat-Mega) assumes that the error terms may follow varying normal distributions at different sites:

$$y_{r,i,j} = \alpha_r + X_{i,j} \cdot \beta_r + \gamma_{r,i} + \delta_{r,i} \cdot \varepsilon_{r,i,j}$$

where $\delta_{r,i}$ are the multiplicative effects of the ith site in the rth ROI.

In addition, it assumes that the additive and multiplicative effects of the sites are not completely independent across ROIs but, rather, they share a common distribution. Such considerations prevent the use of standard linear models, but ComBat uses an empirical Bayes framework to estimate the distribution of the effects of site (Johnson et al., 2007). Once estimated, it derives the additive error terms:

$$\varepsilon_{r,i,j} = \frac{y_{r,i,j} - \alpha_r - X_{i,j} \cdot \beta_r - \gamma_{r,i}}{\delta_{r,i}}$$

These terms allow the derivation of harmonized data:

$$y_{r,i,j}^{ComBat} = \alpha_r + X_{i,j} \cdot \beta_r + \varepsilon_{r,i,j}$$

These simpler data can then be analyzed with standard linear models to estimate the overall difference in cortical thickness between SCZ and CON groups, for each ROI.

2.2. Modifications of the ComBat function

Fortin and colleagues modified the original "combat" function, in the "sva" package for R (Leek et al., 2019), so that it could be applied to imaging data (Fortin et al., 2017). However, Fortin's "combat" function may not be easily applicable to ENIGMA projects as it requires that the dataset has no missing data, which is seldom the case. In addition, it finds the harmonization parameters and applies them to the data within the same function, while some analyses - such as machine learning - require that the parameters are found in a training set and later applied to an independent test set (this is not the case here, but it might be the case in future studies). We further modified the "combat" function to allow for missing data and to separate the fitting and the application of the harmonization.

First, we divided the function into two subfunctions: "combat_fit", which finds the harmonization parameters, and "combat_apply", which applies them to the same or to another set. The "combat_fit" function automatically imputes missing data so that the function can find the harmonization parameters without errors. These imputations are predictions based on linear models of the ROI values by the covariates, separately for each ROI and each site:

$$y_{r,i,j} = \alpha_{r,i} + X_{i,j} \cdot \beta_{r,i}$$

The covariates are the variables introduced into the "combat_fit" function, which in the present study were the diagnosis, age, and sex. The "combat_fit" function also discards ROIs with no variance, which returned errors in the previous "combat" function. Importantly, these imputations are temporary and only aimed to avoid errors during the fitting of the parameters, they are not saved. To apply the parameters, the user must use the "combat_apply" function with the original data, and missing values are not imputed.

The reader may download the adapted ComBat functions for R from http://enigma.ini.usc.edu/wp-content/uploads/combat_for_ENIGMA_sMRI/combat_for_ENIGMA_sMRI.R.

2.3. Collection of data

The data for this paper includes the cortical thickness, surface area and subcortical volumes from 33 sites of the ENIGMA Schizophrenia Working Group (van Erp et al., 2016, 2018; Wong et al., 2019) who shared individual subject level FreeSurfer data for this project. The overall sample included 2897 individuals with a diagnosis of SCZ (mean age 34 years, 34% females) and 3141 CON (mean age 33 years, 49%

Table 2
Description of the overall sample.

	Sample size	Age (SD)	Females	Age of onset (SD)	Duration of illness (SD)	PANSS			SAPS (SD)	SANS (SD)	CDE (SD)
						Total (SD)	Positive (SD)	Negative (SD)			
Patients with schizophrenia	2897	33.9 (12.0)	34.2%	22.8 (7.1)	12.1 (12.5)	60.5 (25.3)	15.5 (6.8)	16.6 (7.8)	20.2 (18.5)	23.0 (16.9)	426 (591)
Healthy controls	3141	33.3 (13.2)	49.0%								

Footnote: CDE: chlorpromazine dose equivalent; PANSS: Positive and Negative Syndrome Scale; SANS: Scale for the Assessment of Negative Symptoms; SAPS: Scale for the Assessment of Positive Symptoms; SD: standard deviation.

females). For SCZ, the mean age of onset was 23 years and their Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987) scores for total/positive/negative symptoms were 61/16/17, respectively. The researchers at each of the sites had collected the data after obtaining participants' written informed consent, with protocols that had been approved by local institutional review boards. We provide a description of the overall sample in Table 2 and a description of the sample from each site in Supplementary Table S1.

All sites had processed the data with FreeSurfer (Fischl, 2012) versions 4.0 to 5.3, except for version 5.2 which was found to produce low intra-class correlations compared to the other versions, and within site all patients and controls were processed using the same FreeSurfer version (van Erp et al., 2016, 2018) according to the ENIGMA protocols, which are available at http://enigma.usc.edu/protocols/imaging-protocols. For cortical ROIs, they involved the estimation of cortical vertex-wise statistics, the extraction of cortical thickness and surface area for 70 Desikan-Killiany (DK) atlas regions (Desikan et al., 2006), and quality checks (van Erp et al., 2018). For subcortical ROIs, they involved the estimation of subcortical volumes and quality checks (van Erp et al., 2016).

2.4. Statistical analyses

We conducted comparisons of MRI data between individuals with SCZ and CON to assess the statistical significance, power and familywise error rate (FWER) using RE-Meta, ME-Mega and ComBat-Mega. We formally tested whether ComBat-Mega increases the statistical significance and power of the differences between individuals with SCZ and CON by attenuating site-effects, using a permutation test and a small-subset strategy respectively. We also used the data of the permutation test to check the FWER.

2.4.1. Comparisons of MRI data between individuals with SCZ and CON

We conducted the RE-Meta in two steps. In the first step, we compared the values of each ROI between SCZ and CON via a standard linear model, with age and sex as covariates, separately for each site. We then converted the difference to a Hedges' *g* and its variance for each site and ROI. In the second step, we conducted a random-effects meta-analysis of the Hedges' *g* of each ROI with the "metafor" package for R (Viechtbauer, 2010), and we corrected the *p*-values for multiple comparisons with the Holm method.

For ME-Mega, we compared the values of each ROI between SCZ and CON via a linear mixed-effects model, with age and sex as covariates and site as a random factor, with the "lme4" and "lmerTest" packages for R (Bates et al., 2015; Kuznetsova et al., 2017). We then divided the difference by the standard deviation (derived from the model) and corrected it for small-sample bias to obtain a Hedges' *g* and its variance, and we corrected the *p*-values for multiple comparisons using the Holm method (Holm, 1979).

Finally, for ComBat-Mega, we first removed the effects of site using the ComBat functions (modelling the effects of diagnosis, age, and sex), and then compared the values of each ROI (e.g., cortical thickness of the frontal pole) between SCZ and CON via a standard linear model, with age and sex as covariates. Note that the ComBat functions use covariates (e.g., age and sex) to better estimate the effects of site, but they do not remove the effects of these covariates; for this reason, we included these covariates in the subsequent linear model. As for ME-Mega, we converted the difference to a Hedges' *g* and its variance, and we corrected the *p*-values for multiple comparisons with the Holm method. Note that we applied a single ComBat harmonization for different types of data (cortical thickness, cortical surface area, and subcortical volume) because we considered that they were related. We also conducted an alternative analysis with a separate harmonization for each type of data.

2.4.2. Comparison of the statistical significance

To test whether ComBat-Mega had improved the statistical significance we used a permutation approach. We followed the Draper-Stoneman procedure, which according to results from a study comparing different algorithms (Winkler et al., 2014), is one of the procedures that best controls the FWER and that can be safely applied here. Note that other algorithms such as Freedman Lane would produce different permuted data for RE-Meta, ME-Mega and ComBat-Mega, which would be problematic in our study because these unwanted differences could confound other potential differences between the methods. Specifically, we randomly permuted the diagnosis among the individuals within each site and repeated all comparison analysis 1000 times.

To show the differences in statistical significance between methods expected by chance, we plotted the histogram of the median difference in the logit-transformed *p*-values between the methods across the permutations (Fig. 1). For example, in one permutation we randomly assigned study participants to patient or control status. We then compared these randomly assigned patients and controls using RE-Meta, ME-Mega and ComBat-Mega. We then calculated differences between logit-transformed p-values of the ComBat-Mega comparison and logit-transformed p-values of the RE-Meta (or ME-Mega) comparisons for each ROI. From these, we only saved the median between logit-transformed p-value difference. Note that this median difference should be very close to zero, given that participant assignment was random, and there should therefore be no patient-control group differences other than by chance. By conducting multiple of these permutations, we were able plot the histogram of the median differences expected by chance alone. Finally, we compared the median difference of the original analysis (with correctly assigned patient and control status) with the histogram of the median differences expected by chance. Only median differences were used in this analysis to simplify the test as doing so avoids the need to correct for multiple comparisons.

We must note that without the logit (or other) transforms, the detection of differences in statistical significance would be too sensitive for large *p*-values and too little sensitive for small *p*-values. For example, if the (non-transformed) *p*-value using one approach was 0.6 and the (non-transformed) *p*-value using another approach was 0.4, the difference in *p*-values would be very large (0.6–0.4 = 0.2) even if the two *p*-values might be considered conceptually very similar, whereas if the (non-transformed) *p*-value using one approach was 0.003 and the (non-transformed) *p*-value using another approach was 0.001, the difference in

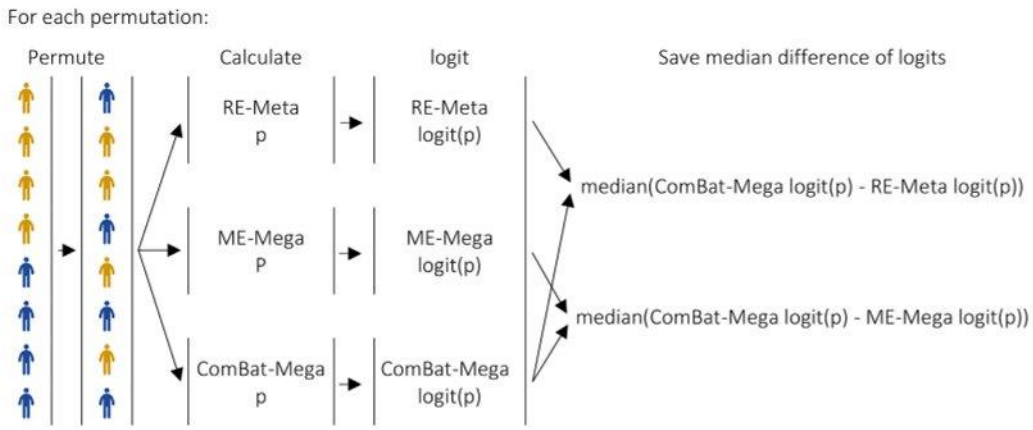


Fig. 1. Steps of each iteration of the permutation test used to compare the statistical significance between random-effects meta-analysis, mixed-effects mega-analysis and ComBat mega-analysis.
Footnote: ComBat-Mega: ComBat mega-analysis; ME-Mega: mixed-effects mega-analysis; RE-Meta: random-effects meta-analysis.

p-values would be very small (0.003–0.001 = 0.002) even if one *p*-value is three times the size of the other. With the logit transform, the *p*-values of the first example would be 0.4 and −0.4, with a difference of 0.8, and the *p*-values of the second example would be −5.8 and −6.9, with a difference of 1.1.

The use of a permutation test implied that both the estimated probability of obtaining the observed median difference in (logit-transform) *p*-values was discrete, i.e., it could only be 0.001, or 0.002, or 0.003,

Table 3
Effect sizes and confidence intervals derived from the ComBat mega-analysis.

		Thickness	Surface area			Thickness	Surface area
Bankssts	L	−0.37 (−0.43,-0.32)	−0.2 (−0.25,-0.14)	Postcentral	L	−0.3 (−0.36,-0.25)	−0.24 (−0.29,-0.19)
	R	−0.39 (−0.44,-0.33)	−0.2 (−0.26,-0.15)		R	−0.28 (−0.33,-0.23)	−0.22 (−0.27,-0.16)
Caudal anterior cingulate	L	−0.12 (−0.18,-0.07)	−0.16 (−0.21,-0.11)	Posterior cingulate	L	−0.24 (−0.3,-0.19)	−0.13 (−0.19,-0.08)
	R	−0.15 (−0.2,-0.1)	−0.2 (−0.26,-0.15)		R	−0.28 (−0.34,-0.23)	−0.18 (−0.23,-0.13)
Caudal middle frontal	L	−0.36 (−0.41,-0.3)	−0.18 (−0.23,-0.13)	Precentral	L	−0.38 (−0.43,-0.32)	−0.19 (−0.24,-0.14)
	R	−0.33 (−0.38,-0.27)	−0.18 (−0.23,-0.13)		R	−0.38 (−0.43,-0.32)	−0.2 (−0.26,-0.15)
Cuneus	L	−0.15 (−0.21,-0.1)	−0.19 (−0.24,-0.13)	Precuneus	L	−0.31 (−0.36,-0.25)	−0.17 (−0.23,-0.12)
	R	−0.19 (−0.24,-0.14)	−0.14 (−0.19,-0.09)		R	−0.34 (−0.4,-0.29)	−0.17 (−0.22,-0.11)
Entorhinal	L	−0.11 (−0.17,-0.06)	−0.16 (−0.21,-0.1)	Rostral anterior cingulate	L	−0.11 (−0.17,-0.06)	−0.17 (−0.22,-0.12)
	R	−0.07 (−0.12,-0.01)	−0.1 (−0.16,-0.05)		R	−0.13 (−0.18,-0.08)	−0.18 (−0.24,-0.13)
Frontal pole	L	−0.19 (−0.24,-0.13)	−0.18 (−0.23,-0.13)	Rostral middle frontal	L	−0.26 (−0.32,-0.21)	−0.24 (−0.29,-0.18)
	R	−0.2 (−0.25,-0.14)	−0.09 (−0.15,-0.04)		R	−0.3 (−0.35,-0.24)	−0.21 (−0.26,-0.16)
Fusiform	L	−0.44 (−0.49,-0.38)	−0.22 (−0.27,-0.17)	Superior frontal	L	−0.33 (−0.38,-0.28)	−0.24 (−0.3,-0.19)
	R	−0.45 (−0.5,-0.39)	−0.26 (−0.32,-0.21)		R	−0.35 (−0.41,-0.3)	−0.24 (−0.29,-0.18)
Inferior parietal	L	−0.41 (−0.47,-0.36)	−0.22 (−0.27,-0.16)	Superior parietal	L	−0.28 (−0.33,-0.23)	−0.2 (−0.25,-0.14)
	R	−0.38 (−0.43,-0.33)	−0.22 (−0.28,-0.17)		R	−0.29 (−0.35,-0.24)	−0.22 (−0.27,-0.17)
Inferior temporal	L	−0.39 (−0.44,-0.33)	−0.25 (−0.31,-0.2)	Superior temporal	L	−0.36 (−0.41,-0.3)	−0.22 (−0.27,-0.17)
	R	−0.34 (−0.39,-0.29)	−0.22 (−0.27,-0.16)		R	−0.38 (−0.43,-0.32)	−0.23 (−0.29,-0.18)
Insula	L	−0.37 (−0.43,-0.32)	−0.17 (−0.22,-0.11)	Supramarginal	L	−0.42 (−0.47,-0.36)	−0.17 (−0.23,-0.12)
	R	−0.37 (−0.42,-0.32)	−0.13 (−0.18,-0.07)		R	−0.39 (−0.44,-0.34)	−0.19 (−0.25,-0.14)
Isthmus cingulate	L	−0.25 (−0.3,-0.2)	−0.06 (−0.12,-0.01)	Temporal pole	L	−0.17 (−0.22,-0.12)	−0.09 (−0.14,-0.03)
	R	−0.25 (−0.3,-0.2)	−0.09 (−0.14,-0.04)		R	−0.17 (−0.22,-0.11)	−0.07 (−0.12,-0.01)
Lateral occipital	L	−0.27 (−0.33,-0.22)	−0.19 (−0.24,-0.13)	Transverse temporal	L	−0.26 (−0.31,-0.2)	−0.15 (−0.21,-0.1)
	R	−0.29 (−0.35,-0.24)	−0.18 (−0.24,-0.13)		R	−0.29 (−0.34,-0.23)	−0.19 (−0.24,-0.14)
Lateral orbitofrontal	L	−0.3 (−0.35,-0.24)	−0.2 (−0.25,-0.14)				
	R	−0.34 (−0.39,-0.29)	−0.19 (−0.24,-0.14)				
Lingual	L	−0.3 (−0.35,-0.24)	−0.21 (−0.26,-0.16)				
	R	−0.32 (−0.37,-0.27)	−0.18 (−0.23,-0.13)			Volume	
Medial orbitofrontal	L	−0.2 (−0.25,-0.14)	−0.19 (−0.25,-0.14)	Accumbens	L	−0.06 (−0.11,-0.01)	
	R	−0.25 (−0.31,-0.2)	−0.19 (−0.25,-0.14)		R	−0.14 (−0.19,-0.09)	
Middle temporal	L	−0.38 (−0.44,-0.33)	−0.24 (−0.3,-0.19)	Amygdala	L	−0.25 (−0.3,-0.2)	
	R	−0.36 (−0.41,-0.3)	−0.26 (−0.31,-0.2)		R	−0.24 (−0.3,-0.19)	
Paracentral	L	−0.33 (−0.38,-0.27)	−0.11 (−0.17,-0.06)	Caudate	L	0.03 (−0.03,0.08)	
	R	−0.31 (−0.37,-0.26)	−0.12 (−0.18,-0.07)		R	0.03 (−0.02,0.08)	
Parahippocampal	L	−0.21 (−0.26,-0.15)	−0.12 (−0.17,-0.06)	Hippocampus	L	−0.43 (−0.48,-0.38)	
	R	−0.21 (−0.26,-0.16)	−0.19 (−0.25,-0.14)		R	−0.42 (−0.48,-0.37)	
Pars opercularis	L	−0.36 (−0.42,-0.31)	−0.18 (−0.23,-0.13)	Lateral Ventricle	L	0.25 (0.19,0.3)	
	R	−0.38 (−0.44,-0.33)	−0.2 (−0.26,-0.15)		R	0.2 (0.15,0.26)	
Pars orbitalis	L	−0.31 (−0.36,-0.26)	−0.21 (−0.26,-0.15)	Pallidum	L	0.28 (0.23,0.33)	
	R	−0.3 (−0.35,-0.25)	−0.17 (−0.23,-0.12)		R	0.19 (0.14,0.24)	
Pars triangularis	L	−0.29 (−0.34,-0.23)	−0.18 (−0.23,-0.12)	Putamen	L	0.09 (0.04,0.15)	
	R	−0.36 (−0.41,-0.3)	−0.16 (−0.22,-0.11)		R	0.1 (0.05,0.15)	
Pericalcarine	L	0 (−0.06,0.05)	−0.14 (−0.19,-0.08)	Thalamus	L	−0.33 (−0.39,-0.28)	
	R	−0.06 (−0.11,0)	−0.09 (−0.15,-0.04)		R	−0.35 (−0.4,-0.29)	

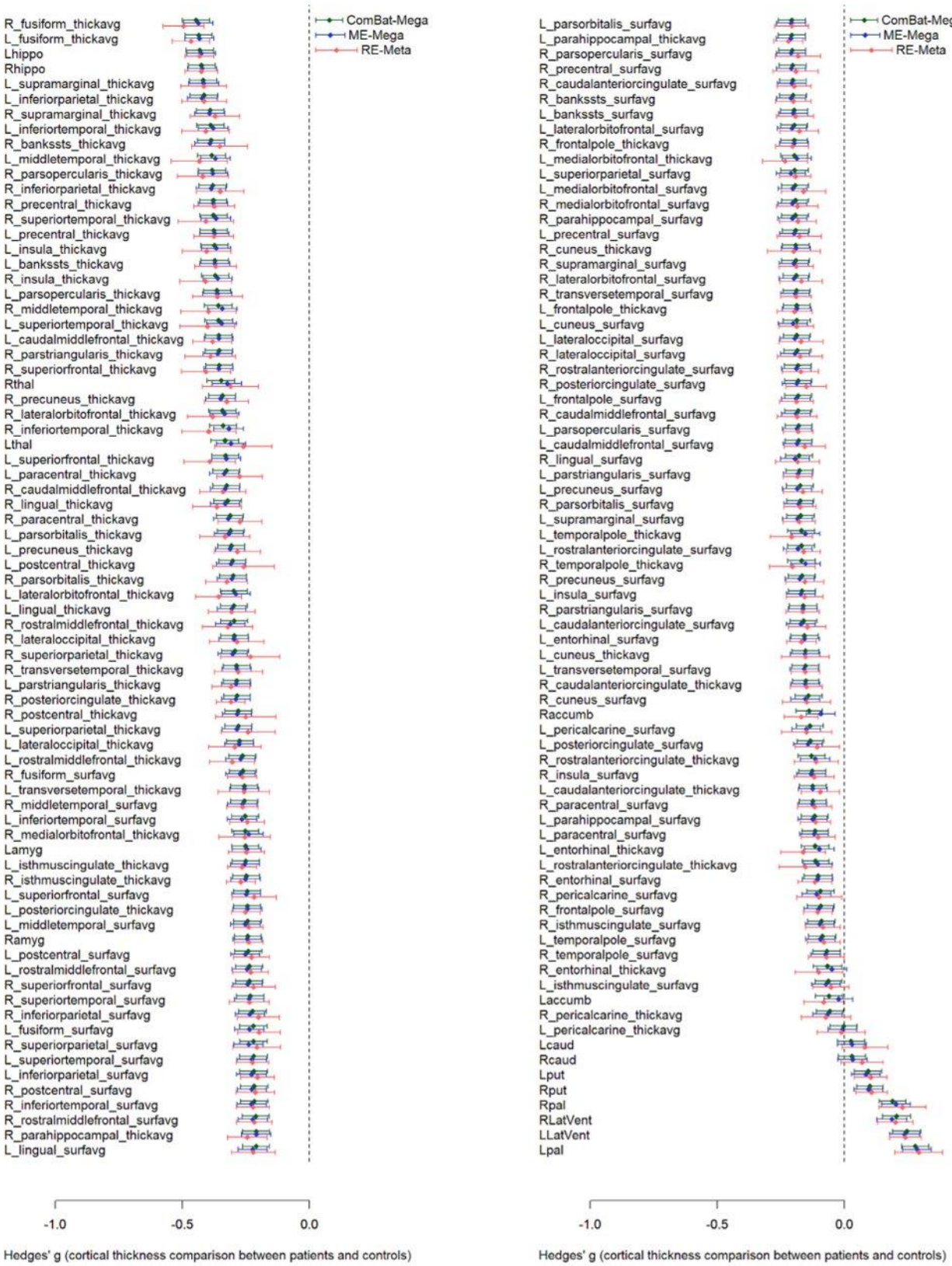


Fig. 2. Forest plot for random-effect meta-analysis (light red), mixed-effects mega-analysis (blue) and ComBat mega-analysis (dark green).
Footnote: The width of the confidence intervals in the legend corresponds to the mean width of the confidence intervals across the brain. ComBat-Mega: ComBat mega-analysis; ME-Mega: mixed-effects mega-analysis; RE-Meta: random-effects meta-analysis.

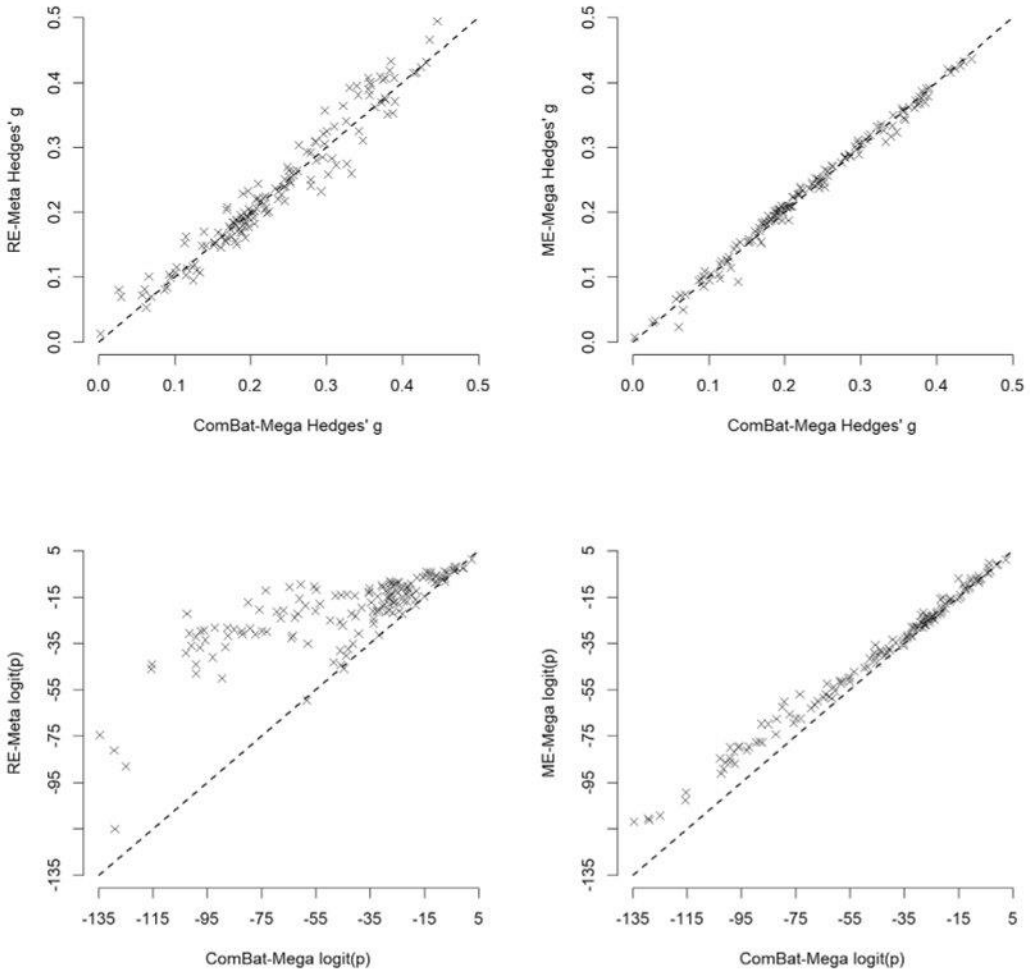


Fig. 3. Hedges' *g* and *p*-values of random-effect meta-analysis, mixed-effects mega-analysis and ComBat mega-analysis in the comparison of ENIGMA brain data between 2897 patients with schizophrenia and 3141 healthy controls.
Footnote: Each cross represents an ROI. ComBat-Mega: ComBat mega-analysis; ME-Mega: mixed-effects mega-analysis; RE-Meta: random-effects meta-analysis. The top plots show that ComBat-Mega effect sizes are similar to RE-Meta and ME-Mega effect sizes, as crosses are mostly distributed around the diagonal lines. The bottom plots show that ComBat-Mega *p*-values are substantially smaller than RE-Meta *p*-values (crosses are clearly above the diagonal line), and slightly smaller than ME-Mega *p*-values (crosses tend to be slightly above the diagonal line).

etcetera. However, we were only interested in assessing if this estimation was <0.05, for what this level of precision should not pose any problems.

2.4.3. Evaluation of the statistical power

We also tested whether ComBat-Mega increases the statistical power using a small-subset strategy. Specifically, we repeated 500 times the analyses but including each time only a random sample of 10 sites. We then counted the number of times that these analyses using only 10 sites were able to detect differences between SCZ and CON. We only used ROIs in which the differences between SCZ and CON were strongly statistically significant in the main analyses using the 33 sites (FWER<0.001 for RE-Meta, for ME-Mega, and for ComBat-Mega), as we assumed that they have true differences. Finally, we conducted a Wilcoxon signed-ranked test to compare the statistical power across ROIs between ComBat-Mega and RE-Meta, as well as between ComBat-Mega and ME-Mega.

2.4.4. Determination of the empirical FWER

We also used the permutation data created above to check whether the FWER for the three methods were appropriate, i.e., we counted the proportion of permutations in which at least one ROI had a Holm-corrected *p*-value < 0.05. Again, the use of a permutation test implied that the estimated FWER was discrete, but we were only interested in assessing whether it was <0.05.

3. Results

With ComBat-Mega, on average, individuals with a diagnosis of SCZ showed thinner cortex and smaller surface area in nearly all cortical ROIs (Table 3). The only exceptions were the bilateral pericalcarine fissures and right entorhinal cortex (where between-group differences in thickness did not reach statistical significance after correction for multiple comparisons) and the left isthmus of the cingulate and right temporal pole (where between-group differences in surface area did not reach statistical significance after correction for multiple comparisons). The SCZ group also showed, on average, smaller bilateral thalamus, hippocampus, amygdala, and right accumbens volumes, and larger bilateral lateral ventricle, putamen, and pallidum volumes. Smaller left accumbens and larger bilateral caudate volumes were not statistically significant after correction for multiple comparisons.

Results were in the same direction for the RE-Meta and ME-Mega, though RE-Meta did not detect thinner cortex in three ROIs (bilateral rostral anterior cingulate and left caudal anterior cingulate) and smaller surface area in six ROIs (bilateral pericalcarine fissure, left posterior cingulate and temporal pole, and right isthmus cingulate and insula).

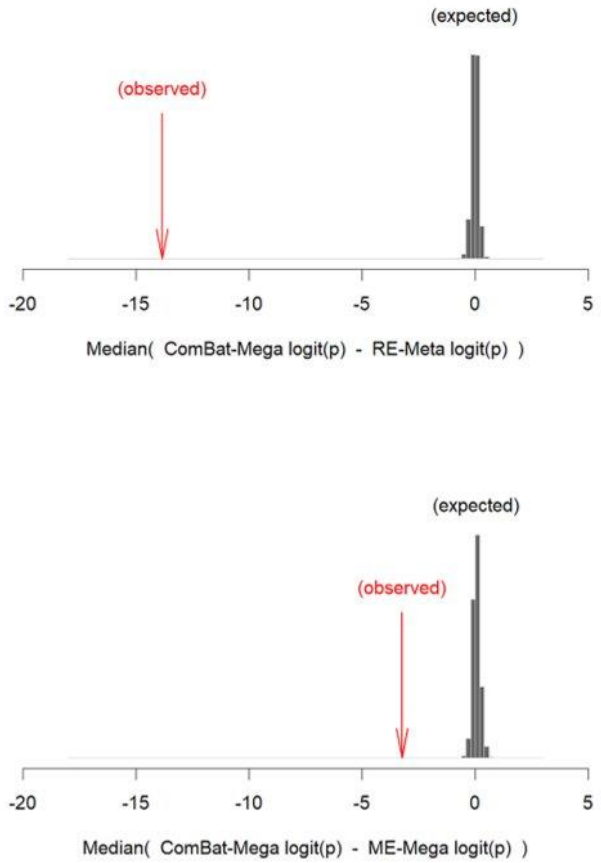


Fig. 4. Median difference between logit-transformed *p*-values derived from ComBat mega-analysis and logit-transformed *p*-values derived from random-effects meta-analysis and mixed-effects mega-analysis in the original data (red) and in the permuted data (histograms).
Footnote: ComBat-Mega: ComBat mega-analysis; ME-Mega: mixed-effects mega-analysis; RE-Meta: random-effects meta-analysis. The histograms (in gray) show the expected ComBat-Mega-related increase of statistical significance by chance, and the arrows (in red) show the actual increase. The latter is clearly larger than that former (negative values mean that ComBat-Mega increases statistical significance).

The Hedges' *g* estimates for the differences were similar across the different analytic methods, but their statistical significance was greater in ComBat-Mega as compared to RE-Meta and ME-Mega (Figs. 2 and 3). The difference in statistical significance was relatively minor when comparing ComBat-Mega to ME-Mega, whereas particularly relevant when comparing ComBat-Mega to RE-Meta (Fig. 3).

The median difference between logit-transformed ComBat-Mega *p*-values and logit-transformed RE-Meta *p*-values in the original data was 13.9. This was substantially larger than any of the median differences in the permuted data (all < 0.61), indicating that the higher statistical significance of ComBat-Mega findings was unlikely due to chance (probability 0.001) (Fig. 4). For the comparison between ComBat-Mega and ME-Mega, the median difference was smaller (3.2), but still unlikely due to chance (all median differences in the permuted data <0.52, probability 0.001).

Interestingly, a plot of the ComBat-Mega-related increase in statistical significance as a function of the intra-site variance/total variance ratio, showed that the increase in statistical significance was larger in those ROIs in which intra-site variance was only ~50–70% of total variance compared to those ROIs in which intra-site variance was ~90–100% of total variance ($p < 0.001$, Fig. 5).

In the evaluation of statistical power using the small-subset strategy, the statistical power was higher for ComBat-Mega (statistical power = 83.5%) than for RE-Meta (statistical power = 53.7%; Wilcoxon *p*-value < 0.001) or ME-Mega (statistical power = 80.4%; Wilcoxon *p*-value < 0.001).

The empirical FWER was ≤ 0.05 for all analytic methods (RE-Meta: 0.024; ME-Mega: 0.027; ComBat-Mega: 0.025).

When we applied the ComBat harmonization separately for cortical thickness data, cortical surface area data and subcortical volume data, we found the same differences with nearly identical Hedges' *g* (Supplementary Figure S1). The statistical significance was minimally lower (median difference between single ComBat logit-transformed *p*-values and separate ComBat logit-transformed *p*-values was 0.1), the statistical power in the small-subset strategy was 83.5%, and the empirical FWER was 0.026.

When we covaried ComBat-Mega by age, sex and ICV, results were similar: The only differences were that the right frontal pole, isthmus of the cingulate and pericalcarine and left parahippocampal and temporal pole decreases in surface area were no longer statistically significant, whereas the left pericalcarine decrease in surface area and the bilateral caudate increases in volume reached statistical significance. Results were again in the same direction for the RE-Meta and ME-Mega, though RE-Meta did not detect statistically significant differences in 36 of the ROIs showing differences with ComBat-Mega, and ME-Mega did not

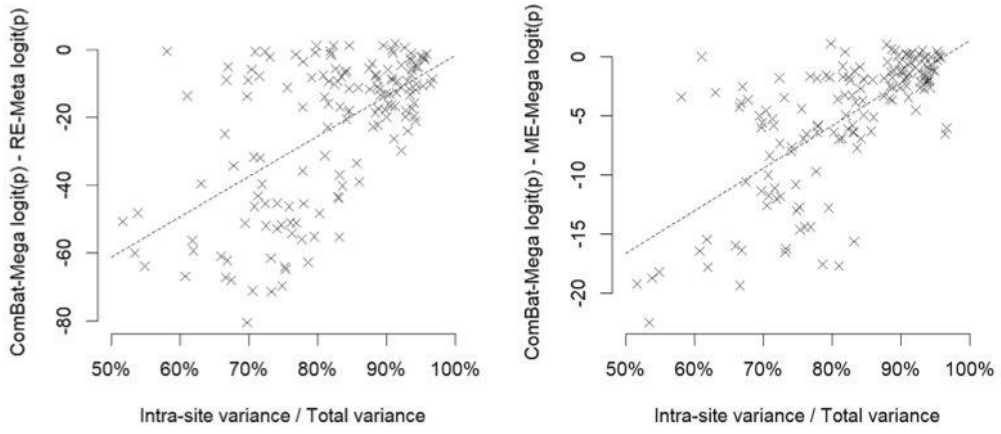


Fig. 5. Relationship between the intra-site variance/total variance ratio and ComBat mega-analysis-related increase of statistical significance.
Footnote: ComBat-Mega: ComBat mega-analysis; ME-Mega: mixed-effects mega-analysis; RE-Meta: random-effects meta-analysis. The ComBat-Mega-related increase of statistical significance (negative values in the Y axis) is larger in regions with lower intra-site variance/variance ratio (around 50–70%).

detect smaller right accumbens volume (and detected smaller surface area in left parahippocampal and right pericalcarine but not in left paracentral and right entorhinal). The Hedges' *g* estimates for the differences were again similar across the different analytic methods, but their statistical significance was again greater in ComBat-Mega as compared to RE-Meta and ME-Mega (Supplementary Figure S2).

4. Discussion

In this study, we analyzed individual subject level data pooled by the ENIGMA Schizophrenia Working Group using three methods to account for the effects of site: random-effects meta-analysis (RE-Meta), linear mixed-effects models (ME-Mega), and ComBat harmonization followed by standard linear models (ComBat-Mega). The results of the comparison between SCZ and CON using ComBat-Mega were similar to the studies already published by the ENIGMA Schizophrenia Working Group: SCZ showed a widespread thinner cortex and smaller surface area (van Erp et al., 2018), smaller hippocampus, amygdala, thalamus and accumbens, and larger lateral ventricles, putamen and pallidum (van Erp et al., 2016) than CON. The results of the same comparison using RE-Meta and ME-Mega were in the same direction and had similar effect sizes, although with a lower statistical significance (i.e. wider confidence intervals, larger *p*-values), especially for RE-Meta. In other words, the use of ComBat increased the statistical significance (i.e., narrower confidence intervals, smaller *p*-values) of the differences between SCZ and CON. This was specially apparent in those ROIs in which intra-site variance was only ~50–70% of total variance. ComBat Mega also showed increased statistical power when we repeated the analyses with fewer sites. All approaches controlled well the FWER, even too strictly probably due to the use of the Holm method, which is more powerful than the Bonferroni method but still conservative (Blakesley et al., 2009). Findings were similar when covarying by ICV.

Based on these findings, we recommend that ENIGMA mega-analysis projects consider applying the ComBat function to reduce the effects of site, followed by standard statistical analysis without including site as a fixed or random effect in the statistical model. To apply ComBat harmonization, we provide easy-to-use functions for R that work even if there are missing data and they can be trained with data from one set and then applied to data from another.

We must note that we conducted these analyses with the three main types of data used in ENIGMA projects: thickness of cortical ROIs, surface area of cortical ROIs, and volumes of subcortical nuclei. However, some ENIGMA projects use other types of data, such as mean fractional anisotropy of white matter tracts, and we have not explored whether the application of ComBat would be beneficial for these projects. Two notions suggest that ComBat should be broadly beneficial. On the one hand, the ComBat algorithm is not specific for a given type of imaging data. Indeed, while it was developed for genomics data (Johnson et al., 2007), we here successfully applied it to three types of ENIGMA imaging data. Moreover, Fortin and colleagues found that ComBat outperforms other harmonization methods for voxel-based fractional anisotropy and mean diffusivity (Fortin et al., 2017), and Yu et al. found similar results for resting-state functional connectivity and network measures (Yu et al., 2018).

While our findings suggest that ComBat harmonization will be useful for most ENIGMA mega-analyses and other multi-site structural imaging work, we suggest caution when combining different types of data. We conducted a single ComBat harmonization for different types of MRI data because we considered that thickness, area, and volume are related, as they are obtained from the same FreeSurfer output of the T1-weighted image and all measure amounts of gray matter. Indeed, an alternative analysis with separate ComBat harmonization for each type of data yielded nearly identical results. However, we do not know whether the application of a single ComBat harmonization on other combinations of data would behave similarly.

Other popular approaches for pooling neuroimaging data are the voxel-based meta-analytic methods, such as Seed-based *d* Mapping (SDM) (Albajes-Eizagirre et al., 2019; Radua et al., 2012) or Activation Likelihood Estimation (ALE) (Eickhoff et al., 2009, 2012). These methods can include imaging studies even if they only report the coordinates of the peaks of the clusters of statistical significance. Therefore, a great advantage of these methods is the exhaustive inclusion of studies. In addition, the analyses are conducted at the voxel level (rather than using ROIs). These methods traditionally tested whether the reported findings tended to converge in a few brain voxels (Albajes-Eizagirre and Radua, 2018), but novel methods are able to directly test whether there are differences – even if they are widespread and do not converge (Albajes-Eizagirre et al., 2019). In view of the results of the present study, one could wonder whether these voxel-based methods should also conduct ComBat mega-analysis instead of meta-analysis. However, to use ComBat they would need access to individual subject level data, which at present are often not available. Another aspect to consider is whether we need SDM or ALE meta-analyses after an ENIGMA ComBat mega-analysis is published. Here, we must remember that SDM and ALE are voxel-based and include virtually all published studies, whereas most ENIGMA studies are ROI-based and include only the data that authors agree to share. Therefore, these different approaches present interesting complementary information.

Our study has some limitations. First, we already stated that we have not explored whether the application of ComBat would be beneficial for projects using other types of data, although several facts suggest that ComBat should be broadly beneficial. Second, we also acknowledged that we do not know whether the application of a single ComBat harmonization on other combinations of data would behave similarly. Third, our analysis is focused on the differences between SCZ and CON, whose distribution is roughly similar across sites. The effects of site and thus the importance of their removal might be larger for conditions with few cases in each site, where pooling data is more beneficial. Fourth, ComBat-Mega addresses some issues but not others, which still need to be investigated, such as site by nuisance confounds. For example, a site with poor quality data may also be a site with a mean age higher than other sites. Future studies addressing these questions could point to methods other than ComBat. Finally, there is a conceptual difference in the effects of site that are modeled in ComBat/ME-Mega and the effects of site that are modeled in RE-Meta. The former effects are in (individual) raw data and refer to site-specific constants that are added to or that multiply the measurements. The latter effects, conversely, are in (group) effect sizes, and are probably a mix of several factors such as site-specific constants that multiply the measurements, heterogeneity in the differences between SCZ and CON, or differences in precision between studies.

To conclude, this paper provides evidence of the superiority of ComBat harmonization over standard mega-analyses and meta-analyses in reducing site-related heterogeneity and thus increase statistical power. We therefore recommend that ENIGMA mega-analysis projects and other multi-site structural imaging work consider applying the ComBat function, which we provide employing easy functions for R. The provided code works with missing data and allows for harmonization of a test set based on the training set (a requirement for machine learning and possibly replication studies). We hope that future ENIGMA mega-analysis projects will improve between-site harmonization using ComBat.

Data accessibility

The adapted ComBat functions for R are available at http://enigma.ini.usc.edu/wp-content/uploads/combat_for_ENIGMA_sMRI/combat_for_ENIGMA_sMRI.R. The data that support the findings of this study may be available on request from the authors of each site participating in the study.

Declaration of competing interest

AB has consulting fees from Biogen and lecture fees from Lundbeck, Otsuka and Janssen. AP has served as a consultant for Boehringer

Ingelheim. AS: Advisory board (DSP), Research grants (CynK, DSP, MTPC, and Ono). CA has been a consultant to or has received honoraria or grants from Acadia, Angelini, Gedeon Richter, Janssen Cilag, Lundbeck, Otsuka, Roche, Sage, Servier, Shire, Schering Plough, Sumitomo Dainippon Pharma, Sunovion and Takeda. CH is faculty member, Lundbeck Psychiatric Institute. CP is on an advisory board for Lundbeck, Australia Pty Ltd and also received honoraria for talks presented at educational meetings organized by Lundbeck. CSW is on an advisory board for Lundbeck, Australia Pty Ltd and in collaboration with Astellas Pharma Inc., Japan. DJS has received research grants or consultancy honoraria from Lundbeck and Sun. EV has received grants and served as a consultant, advisor or CME speaker for the following entities (work unrelated to the topic of this manuscript): AB-Biotics, Abbott, Allergan, Angelini, Dainippon Sumitomo Pharma, Galenica, Janssen, Lundbeck, Novartis, Otsuka, Sage, Sanofi-Aventis, and Takeda. IN has no conflicts of interest to declare. RTS has received consulting fees from Genentech and Roche. SL has received consulting fees and speaking honoraria from Roche, Novartis, Biogen and Merck.

CRediT authorship contribution statement

Joaquim Radua: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing - original draft. **Eduard Vieta:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Russell Shinohara:** Data curation, Formal analysis, Writing - review & editing. **Peter Kochunov:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Yann Quidé:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Melissa J. Green:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Cynthia S. Weickert:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Thomas Weickert:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Jason Bruggemann:** Data curation, Formal analysis, Writing - review & editing. **Tilo Kircher:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Igor Nenadić:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Murray J. Cairns:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Marc Seal:** Data curation, Investigation, Writing - review & editing. **Ulrich Schall:** Data curation, Investigation, Writing - review & editing. **Frans Henskens:** Data curation, Investigation, Writing - review & editing. **Janice M. Fullerton:** Data curation, Investigation, Writing - review & editing. **Bryan Mowry:** Writing - review & editing. **Christos Pantelis:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Rhoshel Lenroot:** Data curation, Investigation, Writing - review & editing. **Vanessa Cropley:** Writing - review & editing. **Carmel Loughland:** Writing - review & editing. **Rodney Scott:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Daniel Wolf:** Data curation, Investigation, Writing - review & editing. **Theodore D. Satterthwaite:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Yunlong Tan:** Data curation, Investigation, Writing - review & editing. **Kang Sim:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Fabrizio Piras:** Data curation, Investigation, Writing - review & editing. **Gianfranco Spalletta:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Nerisa Banaj:** Data curation, Investigation, Writing - review & editing. **Edith Pomarol-Clotet:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Aleix Solanes:** Data curation, Formal analysis, Writing - review & editing. **Anton Albajes-Eizagirre:** Data curation, Formal analysis, Writing - review & editing. **Erick J. Canales-Rodríguez:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Salvador Sarro:** Data curation, Investigation, Writing - review & editing. **Annabella Di Giorgio:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Alessandro Bertolino:** Data curation, Investigation, Writing - review & editing. **Michael Stäblein:** Data curation, Investigation, Writing - review & editing. **Viola Oertel:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Christian Knöchel:** Data curation, Investigation, Writing - review & editing. **Stefan Borgwardt:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Stefan du Plessis:** Writing - review & editing. **Je-Yeon Yun:** Data curation, Formal analysis, Writing - review & editing. **Jun Soo Kwon:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Udo Dannlowski:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Tim Hahn:** Writing - review & editing. **Dominik Grotegerd:** Data curation, Investigation, Writing - review & editing. **Clara Alloza:** Data curation, Formal analysis, Writing - review & editing. **Celso Arango:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Joost Janssen:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Covadonga Díaz-Caneja:** Data curation, Investigation, Writing - review & editing. **Wenhao Jiang:** Writing - review & editing. **Vince Calhoun:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Stefan Ehrlich:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Kun Yang:** Data curation, Formal analysis, Writing - review & editing. **Nicola G. Cascella:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Yoichiro Takayanagi:** Data curation, Investigation, Writing - review & editing. **Akira Sawa:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Alexander Tomyshev:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Irina Lebedeva:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Vasily Kaleda:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Matthias Kirschner:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Cyril Hoschl:** Funding acquisition, Investigation, Writing - review & editing. **David Tomecek:** Data curation, Formal analysis, Writing - review & editing. **Antonin Skoch:** Data curation, Formal analysis, Project administration, Supervision, Writing - review & editing. **Therese van Amelsvoort:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Geor Bakker:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Anthony James:** Data curation, Funding acquisition, Investigation, Writing - review & editing. **Adrian Preda:** Data curation, Investigation, Writing - review & editing. **Andrea Weideman:** Writing - review & editing. **Dan J. Stein:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Fleur Howells:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Anne Uhlmann:** Data curation, Formal analysis, Writing - review & editing. **Henk Temmingh:** Data curation, Investigation, Writing - review & editing. **Carlos López-Jaramillo:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Ana Díaz-Zuluaga:** Data curation, Investigation, Writing - review & editing.

Lydia Fortea: Data curation, Formal analysis, Writing - review & editing. **Eloy Martinez-Heras:** Writing - review & editing. **Elisabeth Solana:** Writing - review & editing. **Sara Llufriu:** Writing - review & editing. **Neda Jahanshad:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Paul Thompson:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Jessica Turner:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Theo van Erp:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - original draft. **David Glahn:** Conceptualization, Funding acquisition, Project administration, Supervision. **Godfrey Pearlson:** Conceptualization, Funding acquisition, Project administration, Supervision. **Axel Krug:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Vaughan Carr:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **Paul Tooney:** Conceptualization, Funding acquisition, Project administration, Supervision. **Gavin Cooper:** Data curation, Investigation, Writing - review & editing. **Paul Rasser:** Data curation, Investigation, Writing - review & editing. **Patricia Michie:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision. **Fude Yang:** Data curation, Investigation. **Federica Piras:** Data curation, Investigation. **Francesca Assogna:** Data curation, Investigation. **Raymond Salvador:** Data curation, Investigation. **Peter McKenna:** Data curation, Investigation. **Aurora Bonvino:** Conceptualization, Funding acquisition, Project administration, Supervision. **Margaret King:** Data curation, Investigation. **Stefan Kaiser:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Dana Nguyen:** Data curation, Investigation. **Julian Pineda-Zapata:** Data curation, Formal analysis.

Acknowledgements

ASRB: The Australian Schizophrenia Research Bank (ASRB) was supported by the National Health and Medical Research Council of Australia (NHMRC) (Enabling Grant, ID 386500), the Pratt Foundation, Ramsay Health Care, the Viertel Charitable Foundation and the Schizophrenia Research Institute. Chief Investigators for ASRB were Carr, V., Schall, U., Scott, R., Jablensky, A., Mowry, B., Michie, P., Catts, S., Henskens, F., Pantelis, C. We thank Loughland, C., the ASRB Manager, and acknowledge the help of Jason Bridge for ASRB database queries. CP was supported by NHMRC Senior Principal Research Fellowships (IDs: 628386 & 1105825); GC was supported by the Schizophrenia Research Institute utilizing infrastructure funding from the New South Wales Ministry of Health and New South Wales Ministry of Trade and Investment (Australia); JMF was supported by NHMRC project grant (1063960) and the Janette Mary O'Neil Research Fellowship; MJG was supported by NHMRC as an R.D. Wright Biomedical Career Development Fellow (1061875). MJC was supported by NHMRC Senior Research Fellowship (1121474).

CASSI: CSW is funded by the NSW Ministry of Health, Office of Health and Medical Research. CSW is a recipient of a National Health and Medical Research Council (Australia) Principal Research Fellowship (PRF) (#1117079).

CIAM: The CIAM study (FMH - PI) was supported by the University Research Committee, University of Cape Town and South African funding bodies National Research Foundation and Medical Research Council.

COBRE: The COBRE dataset and investigators were supported by NIH grants R01EB006841 & P20GM103472, as well as NSF grant 1539067. JT (senior author) and VDC are supported by 5R01MH094524. JMS is supported by R01 AA021771 and P50 AA022534.

EONCKS: This work was supported by a New Partnership for Africa's Development (NEPAD) grant through the Department of Science and Technology of South Africa, the Medical Research Council of South Africa (grant number 65174).

ESO: The ESO study was funded by NPU I – LO1611 and Ministry of Health, Czech Republic – Conceptual Development of Research Organization 00023001 (IKEM).

FIDMAG/Project: This work was supported by the Catalan Government (2017-SGR-1271, 2017-SGR-1365, SLT002/16/00331 and SLT006/17/00357) and several grants from the Instituto de Salud Carlos III and co-funded by European Union (ERDF/ESF, 'Investing in your future'): Miguel Servet Research Contracts (CPII19/00009 to JR, CPII13/00018 to RS and CPII16/00018 to EP-C) and Research Project Grants (PI14/01151, PI14/01148, PI14/00292, PI15/00277, PI15/00283 and PI19/00394).

FOR2107 Marburg: The FOR2107 Marburg study was funded by the German Research Foundation (DFG), Tilo Kircher (speaker FOR2107; DFG grant numbers KI 588/14-1, KI 588/14-2), Axel Krug (KR 3822/5-1, KR 3822/7-2), Igor Nenadic (NE 2254/1-2), Carsten Konrad (KO 4291/3-1).

FOR2107 Muenster: The FOR2107 Muenster study was funded by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to UD) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to UD). TH was supported by grants from the German Research Foundation (DFG grants HA7070/2-2, HA7070/3, HA7070/4).

Frankfurt: MRI was performed at the Frankfurt Brain Imaging Center, supported by the German Research Council (DFG) and the German Ministry for Education and Research (BMBF; Brain Imaging Center Frankfurt/Main, DLR 01GO0203).

GIPSI: This study was supported by Colciencias PRISMA-U.T.

Huilong1 & Huilong2: This study was funded by the National Natural Science Foundation of China (81761128021; 31671145; 81401115; 81401133), Beijing Municipal Science and Technology Commission grant (Z141107002514016) and Beijing Natural Science Foundation(7162087, Beijing Municipal Administration of Hospitals Clinical medicine Development of special funding (XMLX201609; zylx201409).

IGP: This study was funded by Project Grants from the Australian National Health and Medical Research Council of Australia (NHMRC; APP630471 and APP1081603), the Macquarie University's Australian Research Council Centre of Excellence in Cognition and its Disorders (CE110001021).

Johns Hopkins: Supported by National Institutes of Health Grant Nos. MH-092443, MH-094268 (Silvio O. Conte Center), MH-105660, and MH-107730; foundation grants from Stanley, RUSK/S-R, and NARSAD/Brain and Behavior Research Foundation.

Madrid: Supported by the Spanish Ministry of Science, Innovation and Universities, Instituto de Salud Carlos III, co-financed by ERDF Funds from the European Commission, "A way of making Europe", CIBERSAM. Madrid Regional Government (B2017/BMD-3740 AGES-CM-2), European Union Structural Funds and European Union Seventh Framework Program and H2020 Program; Fundación Familia Alonso, Fundación Alicia Koplowitz and Fundación Mutua Madrileña.

MPRC1 & MPRC2: Support was received from NIH grants U01MH108148, 2R01EB015611, R01MH112180, R01DA027680, R01MH085646, P50MH103222 and T32MH067533, a State of Maryland contract (M00B6400091) and NSF grant (1620457).

OLIN: The *Olin* study was supported by NIH grants R37MH43375 and R01MH074797.

Oxford: The Oxford study MRC G0500092.

SLF Rome: Support from the Italian Ministry of Health grants RC-12-13-14-15-16-17-18-19/A.

RSCZ: RSCZ data collection was supported by RFBR 15-06-05758 grant.

SCORE: This study was supported in part by grant 3232BO_119382 from the Swiss National Science Foundation. We thank the FePsy (Frueherkennung von Psychosen; early detection of psychosis) Study Group from the University of Basel, Department of Psychiatry, Switzerland, for the recruitment of the study participants. The FePsy Study was supported in part by grant No. SNF 3200-057216/1, ext./2,

ext./3.

Singapore: This study was supported by research grants from the National Healthcare Group, Singapore (SIG/05004; SIG/11003), and the Singapore Bioimaging Consortium (RP C-009/2006) research grants awarded to KS.

SNUH: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (Grant no. 2013R1A2A1A03071089 and 2017M3C7A1029610).

UCISZ: The UCISZ study was supported by the National Institutes of Mental Health grant number R21MH097196 to TGMvE. UCISZ data were processed by the UCI High Performance Computing cluster supported by Joseph Farran, Harry Mangalam, and Adam Brenner and the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR000153.

UNIBA: The UNIBA study was supported by grant funding from the Italian Ministry of Health (PE-2011-02347951).

UNIMAAS: The study was supported by Dutch Organization for Health Research and Development (ZonMw 91112002) and a personal grant to Thérèse van Amelsvoort (ZonMw-VIDI: 91712394). The data was collected in a clinical trial registered in the Dutch clinical trial registry under ID: NTR5094 (http://www.trialregister.nl).

UPenn: This study was supported by the National Institute of Mental Health grants MH064045, MH 60722, MH019112, MH085096 (DHW), and R01MH112847 (RTS and TDS).

Zurich: This study was supported by the Swiss National Science Foundation (105314_140351 to S.K.). Matthias Kirschner acknowledges support from the National Bank Fellowship (McGill University) and the Swiss National Foundation (P2SKP3_178175).

Research reported in this publication was also supported by the following National Institutes of Health grants: U54 EB020403 to PMT, R01 MH116147, U24 RR21992, R21MH097196, and TR000153 to TGMvE, S10 OD023696 and R01EB015611 to PK, T32 AG058507and 5T32 MH073526 to CRKC, R01 MH117601 to NJ, ENIGMA's NIH Big Data to Knowledge (BD2K) initiative U54 EB020403, ENIGMA Sex Differences R01MH116147, and ENIGMA-COINSTAC: Advanced Worldwide Transdiagnostic Analysis of Valence System Brain Circuits R01MH121246.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116956.

References

Boedhoe, P.S., Schmaal, L., Abe, Y., Ameis, S.H., Arnold, P.D., Batistuzzo, M.C., Benedetti, F., Beucke, J.C., Bollettini, I., Bose, A., Brem, S., Calvo, A., Cheng, Y., Cho, K.I., Dallaspezia, S., Denys, D., Fitzgerald, K.D., Fouche, J.P., Gimenez, M., Gruner, P., Hanna, G.L., Hibar, D.P., Hoexter, M.Q., Hu, H., Huyser, C., Ikari, K., Jahanshad, N., Kathmann, N., Kaufmann, C., Koch, K., Kwon, J.S., Lazaro, L., Liu, Y., Lochner, C., Marsh, R., Martinez-Zalacain, I., Mataix-Cols, D., Menchon, J.M., Minuzzi, L., Nakamae, T., Nakao, T., Narayanaswamy, J.C., Piras, F., Piras, F., Pittenger, C., Reddy, Y.C., Sato, J.R., Simpson, H.B., Soreni, N., Soriano-Mas, C., Spalletta, G., Stevens, M.C., Szeszko, P.R., Tolin, D.F., Venkatasubramanian, G., Walitza, S., Wang, Z., van Wingen, G.A., Xu, J., Xu, X., Yun, J.Y., Zhao, Q., Group, E.O.W., Thompson, P.M., Stein, D.J., van den Heuvel, O.A., 2017. Distinct subcortical volume Alterations in pediatric and adult OCD: a worldwide meta- and mega-analysis. Am. J. Psychiatr. 174, 60–69.

Boedhoe, P.S.W., Heymans, M.W., Schmaal, L., Abe, Y., Alonso, P., Ameis, S.H., Anticevic, A., Arnold, P.D., Batistuzzo, M.C., Benedetti, F., Beucke, J.C., Bollettini, I., Bose, A., Brem, S., Calvo, A., Calvo, R., Cheng, Y., Cho, K.I.K., Ciullo, V., Dallaspezia, S., Denys, D., Feusner, J.D., Fitzgerald, K.D., Fouche, J.P., Fridgeirsson, E.A., Gruner, P., Hanna, G.L., Hibar, D.P., Hoexter, M.Q., Hu, H., Huyser, C., Jahanshad, N., James, A., Kathmann, N., Kaufmann, C., Koch, K., Kwon, J.S., Lazaro, L., Lochner, C., Marsh, R., Martinez-Zalacain, I., Mataix-Cols, D., Menchon, J.M., Minuzzi, L., Morer, A., Nakamae, T., Nakao, T., Narayanaswamy, J.C., Nishida, S., Nurmi, E.L., O'Neill, J., Piacentini, J., Piras, F., Piras, F., Reddy, Y.C.J., Reess, T.J., Sakai, Y., Sato, J.R., Simpson, H.B., Soreni, N., Soriano-Mas, C., Spalletta, G., Stevens, M.C., Szeszko, P.R., Tolin, D.F., van Wingen, G.A., Venkatasubramanian, G., Walitza, S., Wang, Z., Yun, J.Y., Working-Group, E.-O., Thompson, P.M., Stein, D.J., van den Heuvel, O.A., Twisk, J.W.R., 2018. An empirical comparison of meta- and mega-analysis with data from the ENIGMA obsessive-compulsive disorder working group. Front. Neuroinf. 12, 102.

Favre, P., Pauling, M., Stout, J., Hozer, F., Sarrazin, S., Abe, C., Alda, M., Alloza, C., Alonso-Lana, S., Andreassen, O.A., Baune, B.T., Benedetti, F., Busatto, G.F., Canales-Rodriguez, E.J., Caseras, X., Chaim-Avancini, T.M., Ching, C.R.K., Dannlowski, U., Deppe, M., Eyler, L.T., Fatjo-Vilas, M., Foley, S.F., Grotegerd, D., Hajek, T., Haukvik, U.K., Howells, F.M., Jahanshad, N., Kugel, H., Lagerberg, T.V., Lawrie, S.M., Linke, J.O., McIntosh, A., Melloni, E.M.T., Mitchell, P.B., Polosan, M., Pomarol-Clotet, E., Repple, J., Roberts, G., Roos, A., Rosa, P.G.P., Salvador, R., Sarro, S., Schofield, P.R., Serpa, M.H., Sim, K., Stein, D.J., Sussmann, J.E., Temmingh, H.S., Thompson, P.M., Verdolini, N., Vieta, E., Wessa, M., Whalley, H.C., Zanetti, M.V., Leboyer, M., Mangin, J.F., Henry, C., Duchesnay, E., Houenou, J., Group, E.B.D.W., 2019. Widespread white matter microstructural abnormalities in bipolar disorder: evidence from mega- and meta-analyses across 3033 individuals. Neuropsychopharmacology 44, 2285–2293.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., Wright, M.J., Martin, N.G., Agartz, I., Alda, M., Alhusaini, S., Almasy, L., Almeida, J., Alpert, K., Andreasen, N.C., Andreassen, O.A., Apostolova, L.G., Appel, K., Armstrong, N.J., Aribisala, B., Bastin, M.E., Bauer, M., Bearden, C.E., Bergmann, O., Binder, E.B., Blangero, J., Bockholt, H.J., Boen, E., Bois, C., Boomsma, D.I., Booth, T., Bowman, I.J., Bralten, J., Brouwer, R.M., Brunner, H.G., Brohawn, D.G., Buckner, R.L., Buitelaar, J., Bulayeva, K., Bustillo, J.R., Calhoun, V.D., Cannon, D.M., Cantor, R.M., Carless, M.A., Caseras, X., Cavalleri, G.L., Chakravarty, M.M., Chang, K.D., Ching, C.R., Christoforou, A., Cichon, S., Clark, V.P., Conrod, P., Coppola, G., Crespo-Facorro, B., Curran, J.E., Czisch, M., Deary, I.J., de Geus, E.J., den Braber, A., Delvecchio, G., Depondt, C., de Haan, L., de Zubicaray, G.I., Dima, D., Dimitrova, R., Djurovic, S., Dong, H., Donohoe, G., Duggirala, R., Dyer, T.D., Ehrlich, S., Ekman, C.J., Elvsashagen, T., Emsell, L., Erk, S., Espeseth, T., Fagerness, J., Fears, S., Fedko, I., Fernandez, G., Fisher, S.E., Foroud, T., Fox, P.T., Francks, C., Frangou, S., Frey, E.M., Frodl, T., Frouin, V., Garavan, H., Giddaluru, S., Glahn, D.C., Godlewska, B., Goldstein, R.Z., Gollub, R.L., Grabe, H.J., Grimm, O., Gruber, O., Guadalupe, T., Gur, R.E., Gur, R.C., Goring, H.H., Hagenaars, S., Hajek, T., Hall, G.B., Hall, J., Hardy, J., Hartman, C.A., Hass, J., Hatton, S.N., Haukvik, U.K., Hegenscheid, K., Heinz, A., Hickie, I.B., Ho, B.C., Hoehn, D., Hoekstra, P.J., Hollinshead, M., Holmes, A.J., Homuth, G., Hoogman, M., Hong, L.E., Hosten, N., Hottenga, J.J., Hulshoff Pol, H.E., Hwang, K.S., Jack Jr., C.R., Jenkinson, M., Johnston, C., Jonsson, E.G., Kahn, R.S., Kasperaviciute, D., Kelly, S., Kim, S., Kochunov, P., Koenders, L., Kramer, B., Kwok, J.B., Lagopoulos, J., Laje, G., Landen, M., Landman, B.A., Lauriello, J., Lawrie, S.M., Lee, P.H., Le Hellard, S., Lemaitre, H., Leonardo, C.D., Li, C.S., Liberg, B., Liewald, D.C., Liu, X., Lopez, L.M., Loth, E., Lourdusamy, A., Luciano, M., Macciardi, F., Machielsen, M.W., Macqueen, G.M., Malt, U.F., Mandl, R., Manoach, D.S., Martinot, J.L., Matarin, M., Mather, K.A., Mattheisen, M., Mattingsdal, M., Meyer-Lindenberg, A., McDonald, C., McIntosh, A.M., McMahon, F.J., McMahon, K.L., Meisenzahl, E., Melle, I., Milaneschi, Y., Mohnke, S., Montgomery, G.W., Morris, D.W., Moses, E.K., Mueller, B.A., Munoz Maniega, S., Muhleisen, T.W., Muller-Myhsok, B., Mwangi, B., Nauck, M., Nho, K., Nichols, T.E., Nilsson, L.G., Nugent, A.C., Nyberg, L., Olvera, R.L., Oosterlaan, J., Ophoff, R.A., Pandolfo, M., Papalampropoulou-Tsiridou, M., Papmeyer, M., Paus, T., Pausova, Z., Pearlson, G.D., Penninx, B.W., Peterson, C.P., Pfennig, A., Phillips, M., Pike, G.B., Poline, J.B., Potkin, S.G., Putz, B., Ramasamy, A., Rasmussen, J., Rietschel, M., Rijpkema, M., Risacher, S.L., Roffman, J.L., Roiz-Santianez, R., Romanczuk-Seiferth, N., Rose, E.J., Royle, N.A., Rujescu, D., Ryten, M., Sachdev, P.S., Salami, A., Satterthwaite, T.D., Savitz, J., Saykin, A.J., Scanlon, C., Schmaal, L., Schnack, H.G., Schork, A.J., Schulz, S.C., Schur, R., Seidman, L., Shen, L., Shoemaker, J.M., Simmons, A., Sisodiya, S.M., Smith, C., Smoller, J.W., Soares, J.C., Sponheim, S.R., Sprooten, E., Starr, J.M., Steen, V.M., Strakowski, S., Strike, L., Sussmann, J., Samann, P.G., Teumer, A., Toga, A.W., Tordesillas-Gutierrez, D., Trabzuni, D., Trost, S., Turner, J., Van den Heuvel, M., van der Wee, N.J., van Eijk, K., van Erp, T.G., van Haren, N.E., van 't Ent, D., van Tol, M.J., Valdes Hernandez, M.C., Veltman, D.J., Versace, A., Volzke, H., Walker, R., Walter, H., Wang, L., Wardlaw, J.M., Weale, M.E., Weiner, M.W., Wen, W., Westlye, L.T., Whalley, H.C., Whelan, C.D., White, T., Winkler, A.M., Wittfeld, K., Woldehawariat, G., Wolf, C., Zilles, D., Zwiers, M.P., Thalamuthu, A., Schofield, P.R., Freimer, N.B., Lawrence, N.S., Drevets, W., Alzheimer's Disease Neuroimaging Initiative, E.C.I.C.S.Y.S.G, 2014. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav 8, 153–182.

van Erp, T.G., Hibar, D.P., Rasmussen, J.M., Glahn, D.C., Pearlson, G.D., Andreassen, O.A., Agartz, I., Westlye, L.T., Haukvik, U.K., Dale, A.M., Melle, I., Hartberg, C.B., Gruber, O., Kraemer, B., Zilles, D., Donohoe, G., Kelly, S., McDonald, C., Morris, D.W., Cannon, D.M., Corvin, A., Machielsen, M.W., Koenders, L., de Haan, L., Veltman, D.J., Satterthwaite, T.D., Wolf, D.H., Gur, R.C., Gur, R.E., Potkin, S.G., Mathalon, D.H., Mueller, B.A., Preda, A., Macciardi, F., Ehrlich, S., Walton, E., Hass, J., Calhoun, V.D., Bockholt, H.J., Sponheim, S.R., Shoemaker, J.M., van Haren, N.E., Hulshoff Pol, H.E., Ophoff, R.A., Kahn, R.S., Roiz-Santianez, R., Crespo-Facorro, B., Wang, L., Alpert, K.I., Jonsson, E.G., Dimitrova, R., Bois, C., Whalley, H.C., McIntosh, A.M., Lawrie, S.M., Hashimoto, R., Thompson, P.M., Turner, J.A., 2016. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. Mol. Psychiatr. 21, 547–553.

van Erp, T.G.M., Walton, E., Hibar, D.P., Schmaal, L., Jiang, W., Glahn, D.C., Pearlson, G.D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., Yamamori, H., Bustillo, J.R., Clark, V.P., Agartz, I., Mueller, B.A., Cahn, W., de Zwarte, S.M.C.,

Hulshoff Pol, H.E., Kahn, R.S., Ophoff, R.A., van Haren, N.E.M., Andreassen, O.A., Dale, A.M., Doan, N.T., Gurholt, T.P., Hartberg, C.B., Haukvik, U.K., Jorgensen, K.N., Lagerberg, T.V., Melle, I., Westlye, L.T., Gruber, O., Kraemer, B., Richter, A., Zilles, D., Calhoun, V.D., Crespo-Facorro, B., Roiz-Santianez, R., Tordesillas-Gutierrez, D., Loughland, C., Carr, V.J., Catts, S., Cropley, V.L., Fullerton, J.M., Green, M.J., Henskens, F.A., Jablensky, A., Lenroot, R.K., Mowry, B.J., Michie, P.T., Pantelis, C., Quide, Y., Schall, U., Scott, R.J., Cairns, M.J., Seal, M., Tooney, P.A., Rasser, P.E., Cooper, G., Shannon Weickert, C., Weickert, T.W., Morris, D.W., Hong, E., Kochunov, P., Beard, L.M., Gur, R.E., Gur, R.C., Satterthwaite, T.D., Wolf, D.H., Belger, A., Brown, G.G., Ford, J.M., Macciardi, F., Mathalon, D.H., O'Leary, D.S., Potkin, S.G., Preda, A., Voyvodic, J., Lim, K.O., McEwen, S., Yang, F., Tan, Y., Tan, S., Wang, Z., Fan, F., Chen, J., Xiang, H., Tang, S., Guo, H., Wan, P., Wei, D., Bockholt, H.J., Ehrlich, S., Wolthusen, R.P.F., King, M.D., Shoemaker, J.M., Sponheim, S.R., De Haan, L., Koenders, L., Machielsen, M.W., van Amelsvoort, T., Veltman, D.J., Assogna, F., Banaj, N., de Rossi, P., Iorio, M., Piras, F., Spalletta, G., McKenna, P.J., Pomarol-Clotet, E., Salvador, R., Corvin, A., Donohoe, G., Kelly, S., Whelan, C.D., Dickie, E.W., Rotenberg, D., Voineskos, A.N., Ciufolini, S., Radua, J., Dazzan, P., Murray, R., Reis Marques, T., Simmons, A., Borgwardt, S., Egloff, L., Harrisberger, F., Riecher-Rossler, A., Smieskova, R., Alpert, K.I., Wang, L., Jonsson, E.G., Koops, S., Sommer, I.E.C., Bertolino, A., Bonvino, A., Di Giorgio, A., Neilson, E., Mayer, A.R., Stephen, J.M., Kwon, J.S., Yun, J.Y., Cannon, D.M., McDonald, C., Lebedeva, I., Tomyshev, A.S., Akhadov, T., Kaleda, V., Fatouros-Bergman, H., Flyckt, L., Karolinska Schizophrenia, P., Busatto, G.F., Rosa, P.G.P., Serpa, M.H., Zanetti, M.V., Hoschl, C., Skoch, A., Spaniel, F., Tomecek, D., Hagenaars, S.P., McIntosh, A.M., Whalley, H.C., Lawrie, S.M., Knochel, C., Oertel-Knochel, V., Stablein, M., Howells, F.M., Stein, D.J., Temmingh, H.S., Uhlmann, A., Lopez-Jaramillo, C., Dima, D., McMahon, A., Faskowitz, J.I., Gutman, B.A., Jahanshad, N., Thompson, P.M., Turner, J.A., 2018. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing Neuro imaging Genetics through meta analysis (ENIGMA) consortium. Biol. Psychiatr. 84, 644–654.

van Rooij, D., Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., Busatto, G.F., Calderoni, S., Daly, E., Deruelle, C., Di Martino, A., Dinstein, I., Duran, F.L.S., Durston, S., Ecker, C., Fair, D., Fedor, J., Fitzgerald, J., Freitag, C.M., Gallagher, L., Gori, I., Haar, S., Hoekstra, L., Jahanshad, N., Jalbrzikowski, M., Janssen, J., Lerch, J., Luna, B., Martinho, M.M., McGrath, J., Muratori, F., Murphy, C.M., Murphy, D.G.M., O'Hearn, K., Oranje, B., Parellada, M., Retico, A., Rosa, P., Rubia, K., Shook, D., Taylor, M., Thompson, P.M., Tosetti, M., Wallace, G.L., Zhou, F., Buitelaar, J.K., 2018. Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD working group. Am. J. Psychiatr. 175, 359–369.

Wong, T.Y., Radua, J., Pomarol-Clotet, E., Salvador, R., Albajes-Eizagirre, A., Solanes, A., Canales-Rodriguez, E.J., Guerrero-Pedraza, A., Sarro, S., Kircher, T., Nenadic, I., Krug, A., Grotegerd, D., Dannlowski, U., Borgwardt, S., Riecher-Rossler, A., Schmidt, A., Andreou, C., Huber, C.G., Turner, J., Calhoun, V., Jiang, W., Clark, S., Walton, E., Spalletta, G., Banaj, N., Piras, F., Ciullo, V., Vecchio, D., Lebedeva, I., Tomyshev, A.S., Kaleda, V., Klushnik, T., Filho, G.B., Zanetti, M.V., Serpa, M.H., Penteado Rosa, P.G., Hashimoto, R., Fukunaga, M., Richter, A., Kramer, B., Gruber, O., Voineskos, A.N., Dickie, E.W., Tomecek, D., Skoch, A., Spaniel, F., Hoschl, C., Bertolino, A., Bonvino, A., Di Giorgio, A., Holleran, L., Ciufolini, S., Marques, T.R., Dazzan, P., Murray, R., Lamsma, J., Cahn, W., van Haren, N., Diaz-Zuluaga, A.M., Pineda-Zapata, J.A., Vargas, C., Lopez-Jaramillo, C., van Erp, T.G.M., Gur, R.C., Nickl-Jockschat, T., 2019. An overlapping pattern of cerebral cortical thinning is associated with both positive symptoms and aggression in schizophrenia via the ENIGMA consortium. Psychol. Med. 1–12.

Albajes-Eizagirre, A., Radua, J., 2018. What do results from coordinate-based meta-analyses tell us? Neuroimage 176, 550–553.

Albajes-Eizagirre, A., Solanes, A., Vieta, E., Radua, J., 2019. Voxel-based meta-analysis via permutation of subject images (PSI): theory and implementation for SDM. Neuroimage 186, 174–184.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67, 1–48.

Blakesley, R.E., Mazumdar, S., Dew, M.A., Houck, P.R., Tang, G., Reynolds 3rd, C.F., Butters, M.A., 2009. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology 23, 255–264.

Chen, B., Benedetti, A., 2017. Quantifying heterogeneity in individual participant data meta-analysis with binary outcomes. Syst. Rev. 6, 243.

Chepkoech, J.L., Walhovd, K.B., Grydeland, H., Fjell, A.M., Alzheimer's Disease Neuroimaging, I., 2016. Effects of change in FreeSurfer version on classification accuracy of patients with Alzheimer's disease and mild cognitive impairment. Hum. Brain Mapp. 37, 1831–1841.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9, 179–194.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980.

Eickhoff, S.B., Laird, A.R., Grefkes, C., Wang, L.E., Zilles, K., Fox, P.T., 2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. Hum. Brain Mapp. 30, 2907–2926.

Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. Neuroimage 59, 2349–2361.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195–207.

Fortin, J.P., Parker, D., Tunc, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.

Gronenschild, E.H., Habets, P., Jacobs, H.I., Mengelers, R., Rozendaal, N., van Os, J., Marcelis, M., 2012. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One 7, e38234.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Huber, G., 1957. Pneumencephalographische und psychopathologische Bilder bei endogenen Psychosen. Springer, Berlin, Heidelberg.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Software 82, 1–26.

Leek, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., Storey, J.D., Zhang, Y., Torres, L.C., 2019. Sva: Surrogate Variable Analysis. R package.

Radua, J., Mataix-Cols, D., 2012. Meta-analytic methods for neuroimaging data explained. Biol. Mood Anxiety Disord. 2, 6.

Radua, J., Mataix-Cols, D., Phillips, M.L., El-Hage, W., Kronhaus, D.M., Cardoner, N., Surguladze, S., 2012. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. Eur. Psychiatr. 27, 605–611.

Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C., Avants, B.B., 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 99, 166–179.

Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Software 36, 1–48.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.

Yu, M., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., Sheline, Y.I., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp. 39, 4213–4227.