A framework for confounder considerations in AI-driven precision medicine

Vera Komeyer[1,2,3], Prof. Dr. Simon B. Eickhoff[1,2], Prof. Dr. Christian Grefkes[4,5,6], Dr. Kaustubh R. Patil[1,2], Dr. Federico Raimondo[1,2*]*

[1]Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Juelich, Juelich, Germany

[2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Duesseldorf, Duesseldorf, Germany

[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Duesseldorf, Duesseldorf, Germany

[4]Department of Neurology, University Hospital Frankfurt, Goethe University Frankfurt, Frankfurt (Main), Germany

[5]Department of Neurology, University Hospital Cologne and Medical Faculty, University of Cologne, Cologne, Germany

[6]Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Research Centre Juelich, Juelich, Germany

* Correspondence to Vera Komeyer (v.komeyer@fz-juelich.de) and Federico Raimondo (f.raimondo@fz-juelich.de)

Abstract

Introduction Artificial intelligence holds promise for individualized medicine. Yet, transitioning models from prototyping to clinical applications poses challenges, with confounders being a significant hurdle. We introduce a two-dimensional confounder framework (*Confound Continuum*), integrating a statistical dimension with a biomedical perspective. Informed and context-sensitive confounder decisions are indispensable for accurate model building, rigorous evaluation and valid interpretation.

Methods Using prediction of hand grip strength (HGS) from neuroimaging-derived features in a large sample as an example task, we develop a conceptual framework for confounder considerations and integrate it with an exemplary statistical investigation of 130 candidate confounders. We underline the necessity for conceptual considerations by predicting HGS with varying confound removal scenarios, neuroimaging derived features and machine learning algorithms. We use the confounders alone as features or together with grey matter volume to dissect the contribution of the two signal sources.

Results The conceptual confounder framework distinguishes between *high-performance* models and *pure link* models that aim to deepen our understanding of feature-target relationships. The biological attributes of different confounders can overlap to varying degrees with those of the predictive problem space, making the development of *pure link* models increasingly challenging with greater overlap. The degree of biological overlap allows to sort potential confounders on a conceptual *Confound Continuum*. This conceptual continuum complements statistical investigations with biomedical domain-knowledge, represented as an orthogonal two-dimensional grid.

Exemplary HGS predictions highlighted the substantial impact of confounders on predictive performance. In contrast, choice of features or learning algorithms had considerably smaller influences. Notably, models using confounders as features often outperformed models relying solely on neuroimaging features.

Conclusion Our study provides a confounder framework that combines a statistical perspective on confounders and a biomedical perspective. It stresses the importance of domain expertise in predictive modelling for critical and deliberate interpretation and employment of predictive models in biomedical applications and research.

Short description

The paper explores the challenges of transitioning predictive models from scientific prototyping to clinical use, with a focus on the significant impact of confounders. Using the example of predicting hand grip strength in the UK Biobank, the study introduces a framework that integrates statistical and biomedical perspectives on confounders, emphasizing the vital role of informed confounder decisions for accurate model development, evaluation and interpretation.

1. Confounders in precision medicine

Artificial intelligence (AI) holds promise for personalized medicine and is increasingly employed in biomedical research and applications. Machine Learning (ML) workflows use large, high-dimensional and multimodal data to arrive at predictive models to identify biomarkers of health and disease or to aid in diagnosis, prognosis and treatment choice, targeted to individuals[1–3]. For instance, deep learning-based models showed promising results for improved cancer diagnosis, subtyping and staging[4]. Beyond cancer, (chronic) inflammatory diseases stand as significant global contributors to mortality. AI has proven promising in enhancing inflammatory disease risk prediction and facilitating personalized early interventions[5]. In the field of psychiatry, predictive modelling with neuroimaging data has demonstrated the potential to outperform DSM/ICD-based diagnoses[6]. However, translation of promising models to real-world clinical applications still remains challenging, sometimes referred to as AI chasm[7–10]. The AI chasm stems from unreliable predictions[11–14], challenges with reproducibility and replicability, non-interpretability[8], and limited generalizability[15] of models (for further challenges see e.g. [3,7,12,16,17]). Confounding effects contribute significantly to these concerns through misleading predictions and interpretations, thereby exacerbating the AI chasm[18–20].

In a predictive modelling context, confounders are variables that correlate with features and targets, but are not of primary interest or may even introduce misleading associations[21,22] (see e.g.[22–28] for in-depth technical elaborations). Confounders can influence predictions, especially when they carry a strong signal about the target. For example, in a neuroimaging context, a model predicting hand grip strength (HGS) from neuroimaging derived features could be primarily driven by sex, i.e. men on average being stronger than women. Other classical examples of confounders include measurement artifacts[27,29–31], site effects[32], demographics[33–35], or lifestyle factors[36].

It is essential to deal with confounding effects to obtain models that give valid scientific insights and models that can be deployed in clinical practice. Established tools to control for confounders at the level of study design, such as randomized control trials, restriction or matching[23], may not be feasible in observational data[6,20,22,37]. Consequently, post-hoc statistical approaches, such as (linear) confounder regression are commonly used[18,19,24,27,38–40]. Alternatively, the contribution of confounders can be quantified by including them as predictors[18,27,41].

In many biomedical disciplines it is common to correct for a conventionally established set of confounders[18,42,43]. While for instance in the field of genetics it is common to adjust for a broader set of confounders, in neuroimaging studies sex and age are most prevalently considered[12,44]. This reliance on convention, however, risks overlooking other potential confounders. Overlooking confounders or insufficient removal of their signal contributions can lead to overestimated effects because predictions are driven by confounding signals rather than the actual signal of interest[18]. Conversely, removal of too many confounders can eliminate signal of interest and lead to unstable models[32,45,46]. Adjusting for a variable that is actually a consequence of the features (i.e. not a real confounder) may even induce a non-existent association (Berkson's paradox)[18,47–49]. Overall, generic conventional and non-contextual confound removal (too many or too little) can result in suboptimal models[19,22]. Consequently, it is important to identify confounders that align with the goals of the modelling task at hand[18,19,22]. Furthermore, even if a suitable set of context-specific confounders is identified, particularly in a research context it often remains unclear whether a "vanilla" model with no confound removal or a confounder adjusted model should be preferred. Taken together, suboptimal treatment of confounding contributes to the challenges of transitioning models from development to clinical applications, aggravating the AI chasm.

The goal of this paper is to emphasize a better understanding of confounders for a given research endeavour. We elaborate on the necessity of acknowledging domain expertise and biomedical knowledge to form a biomedical dimension of confounder considerations. We introduce a two-dimensional (2D) grid (*Confound Continuum*), of which the horizontal axis acknowledges the degree of biomedical impact of a confounder on a predictive problem, while the vertical axis evaluates the statistical impact. Adopting such an integrated perspective of statistical and biomedical confounder considerations fosters informed, context-sensitive decisions on confounders as an indispensable step towards accurate and valid model development

and interpretation. Our aim is to encourage critical and deliberate employment of AI, both for medical applications and biomedical research.

2. Defining the context of confound removal

2.1. The statistical context of an exemplary GMV-HGS prediction

We illustrate concepts with the example prediction of hand grip strength (HGS) from grey matter volume (GMV) features in the UK Biobank[50]. HGS is an ideal target variable for this demonstration. It is reliable[51,52] and eliminates further complexities associated with latent target measures such as intelligence or executive functioning scores. Additionally, HGS is an objective and cost-effective assessment commonly used in clinical settings[53].

Commonly, the relevance of a set of candidate confounders is determined by assessing their statistical association with the data. Variables with strong associations or high shared variance are considered as confounders in the predictive analysis. Understanding such associations is crucial because removing confounders without shared signal may inadvertently introduce confounder information into features or target[54]. Conversely, removing confounders with high shared variance may enhance the signal-to-noise ratio of the feature-target relationship.

Mimicking such a statistical approach, we exemplarily correlated 130 candidate confounders from the UKB with both HGS and GMV (**Figure 1**). The correlations revealed that mostly body composition measures, sex and respiratory variables were associated with either the target HGS or the GMV features. Variables such as “length of the working week in the main job”, “systolic blood pressure”, “age” and “bone density” exhibited medium to small correlations with HGS or GMV. For a more comprehensive statistical investigation of confounders in the UKB see e.g.[18].

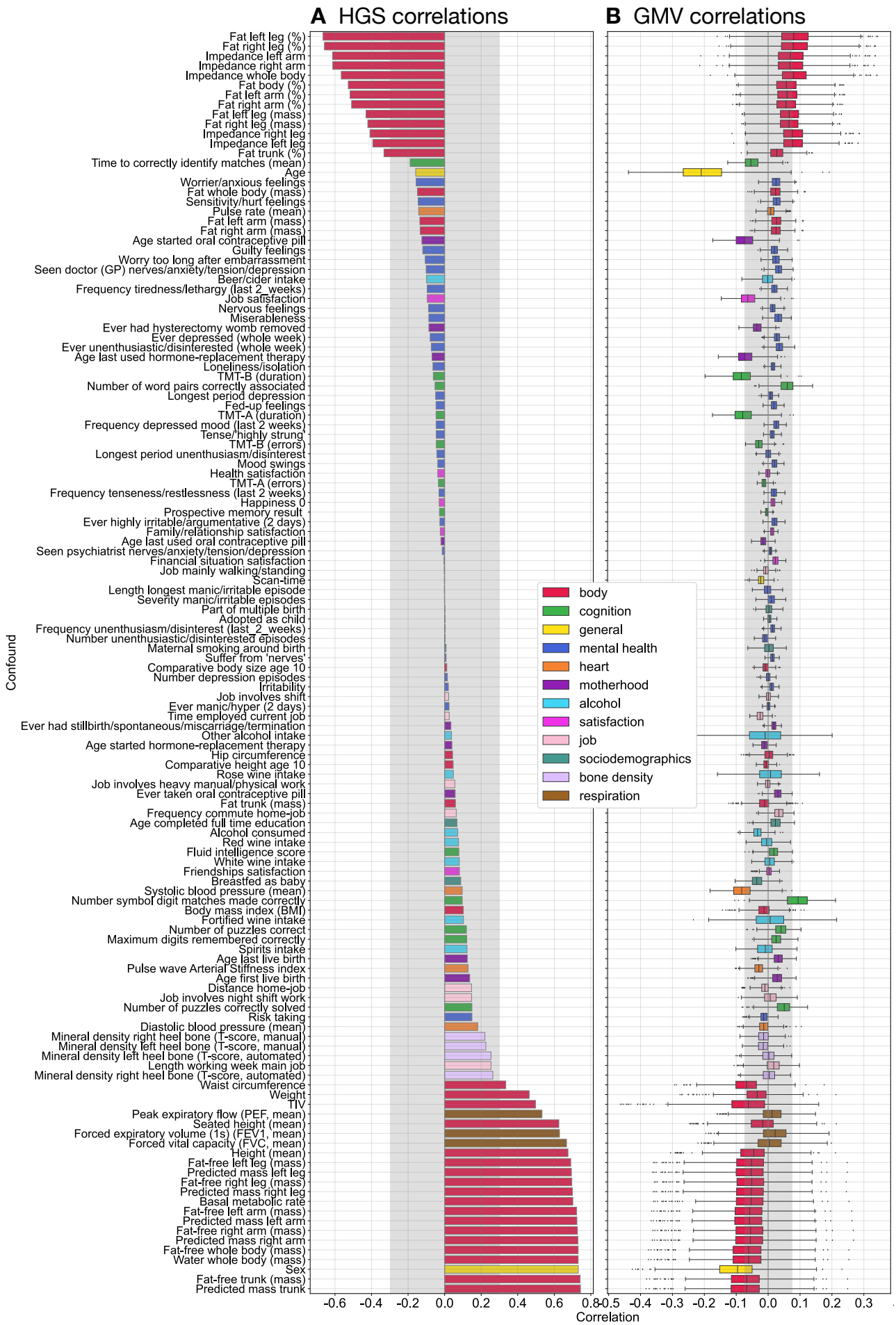
A HGS correlations
B GMV correlations
Confound
Fat left leg (%)
Fat right leg (%)
Impedance left arm
Impedance right arm
Impedance whole body
Fat body (%)
Fat left arm (%)
Fat right arm (%)
Fat left leg (mass)
Fat right leg (mass)
Impedance right leg
Impedance left leg
Fat trunk (%)
Time to correctly identify matches (mean)
Age
Worrier/anxious feelings
Fat whole body (mass)
Sensitivity/hurt feelings
Pulse rate (mean)
Fat left arm (mass)
Fat right arm (mass)
Age started oral contraceptive pill
Guilty feelings
Worry too long after embarrassment
Seen doctor (GP) nerves/anxiety/tension/depression
Beer/cider intake
Frequency tiredness/lethargy (last 2_weeks)
Job satisfaction
Nervous feelings
Miserableness
Ever had hysterectomy womb removed
Ever depressed (whole week)
Ever unenthusiastic/disinterested (whole week)
Age last used hormone-replacement therapy
Loneliness/isolation
TMT-B (duration)
Number of word pairs correctly associated
Longest period depression
Fed-up feelings
TMT-A (duration)
Frequency depressed mood (last 2 weeks)
Tense/highly strung
TMT-B (errors)
Longest period unenthusiasm/disinterest
Mood swings
Health satisfaction
TMT-A (errors)
Frequency tenseness/restlessness (last 2 weeks)
Happiness 0
Prospective memory result
Ever highly irritable/argumentative (2 days)
Family/relationship satisfaction
Age last used oral contraceptive pill
Seen psychiatrist nerves/anxiety/tension/depression
Financial situation satisfaction
Job mainly walking/standing
Scan-time
Length longest manic/irritable episode
Severity manic/irritable episodes
Part of multiple birth
Adopted as child
Frequency unenthusiasm/disinterest (last_2_weeks)
Number unenthusiastic/disinterested episodes
Maternal smoking around birth
Suffer from 'nerves'
Comparative body size age 10
Number depression episodes
Irritability
Job involves shift
Ever manic/hyper (2 days)
Time employed current job
Ever had stillbirth/spontaneous/miscarriage/termination
Other alcohol intake
Age started hormone-replacement therapy
Hip circumference
Comparative height age 10
Rose wine intake
Job involves heavy manual/physical work
Ever taken oral contraceptive pill
Fat trunk (mass)
Frequency commute home-job
Age completed full time education
Alcohol consumed
Red wine intake
Fluid intelligence score
White wine intake
Friendships satisfaction
Breastfed as baby
Systolic blood pressure (mean)
Number symbol digit matches made correctly
Body mass index (BMI)
Fortified wine intake
Number of puzzles correct
Maximum digits remembered correctly
Spirits intake
Age last live birth
Pulse wave Arterial Stiffness index
Age first live birth
Distance home-job
Job involves night shift work
Number of puzzles correctly solved
Risk taking
Diastolic blood pressure (mean)
Mineral density right heel bone (T-score, manual)
Mineral density left heel bone (T-score, manual)
Mineral density left heel bone (T-score, automated)
Length working week main job
Mineral density right heel bone (T-score, automated)
Waist circumference
Weight
TIV
Peak expiratory flow (PEF, mean)
Seated height (mean)
Forced expiratory volume (1s) (FEV1, mean)
Forced vital capacity (FVC, mean)
Height (mean)
Fat-free left leg (mass)
Predicted mass left leg
Fat-free right leg (mass)
Predicted mass right leg
Basal metabolic rate
Fat-free left arm (mass)
Predicted mass left arm
Fat-free right arm (mass)
Predicted mass right arm
Fat-free whole body (mass)
Water whole body (mass)
Sex
Fat-free trunk (mass)
Predicted mass trunk
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8
-0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3
Correlation
body
cognition
general
mental health
heart
motherhood
alcohol
satisfaction
job
sociodemographics
bone density
respiration

Figure 1. Correlations of 130 summary behavioural variables with the exemplary target HGS (**A**) and the features GMV (**B**) that could potentially be considered as confounding variables. The variables were sorted into 12 higher-level categories. Boxplots in B) indicate median (IQR) correlation over GMV parcels. Correlations refer to Pearson's r for continuous confounds, Spearman correlations for ordinal variables and point-biserial correlation coefficients for binary variables.

2.2. *High-performance* versus *pure link* approach motivates a conceptual context of confound removal

To develop unbiased models, beyond statistics, it is crucial to understand the context of a prediction task. We therefore introduce the distinction between a *high-performance* model and a model aimed at investigating a *pure link* as overarching research goal. Both setups require careful consideration of confounders.

The *high-performance* approach aims to achieve accurate predictions by utilizing all available information irrespective of its origin. Here, confounders may even be included as features if they improve model accuracy. It nevertheless remains essential to satisfy the fundamental assumption of predictive modelling that training and testing data are drawn from the same distribution and are independent and identically distributed (iid). Satisfying the iid assumption avoids sampling bias and helps build generalizable models that can apply patterns learned from the training set to unseen testing data. Otherwise, a model may perform well on training but fail on testing data, exacerbating the AI chasm. Differences in training and testing data distribution are sometimes referred to as data distribution shift[55]. For instance in healthcare applications, differences in patient demographics or medical practices between hospitals can cause such a shift. Covariate shift is a specific form thereof, where particularly the distribution of the independent variables (features and/or confounders) changes[56]. To avoid shift-related issues, even in the *high-performance* setting, training and testing data must be comparable in their key characteristics, including their relationship with confounders.

The *pure link* setting aims to deepen our understanding of specific feature-target relations by discovering systematic, biologic mechanisms underlying the feature-target interactions. Such models selectively utilize specific aspects of the available information in the data. Concretely, this approach prioritizes the signal components in the features that hold biomedical meaning to predict the respective outcome, such as a phenotype, behaviour or disease, but aims to exclude encoded information of confounders in the biomedical feature signal (e.g. neuroimaging-derived features). By doing so, it aims to uncover the "pure" biology of the problem space and contribute to a broader comprehension of biomedical mechanisms.

However, achieving such "purity" becomes an idealized goal when dealing with biologically highly linked confounders. To illustrate this challenge, we consider two of the statistically evaluated potential confounding variables for the GMV-HGS prediction task: "Length of working week in the main job" and "sex" (**Figure 2A**). Unlike "length of working week", "sex" significantly overlaps in its biological attributes with those of the GMV-HGS problem space, i.e. "sex" and the problem space have a high "shared biology" (not to be confused with a high shared variance in a statistical sense). From biomedical domain knowledge it is known that sex influences testosterone levels, which, in turn, impact muscle growth and the muscle mass determines HGS. In the *pure link* setup, confound removal is expected to preserve all meaningful connections between GMV and HGS, while eliminating unwanted influence of confounders, expecting to obtain the "pure" biology of the problem space (**Figure 2A** bottom: middle & left). This expectation of "purity" can be fulfilled for non-overlapping variables, such as "length of working week" (which in this extreme would then not be considered as confounder). However, removing highly overlapping variables, such as "sex", results in a new (artificial) set of biological attributes of the GMV-HGS problem space (**Figure 2A** bottom: right, non-circular red outline). This artificial shape is biologically ambiguous and challenging to interpret. Consequently, the more a confounder overlaps in its biology with the problem space, the less its removal can lead to "purity". This problem particularly arises in the biomedical field due to the low-dimensionality and interconnected nature of many biological phenomena and necessitates to acknowledge a conceptual dimension of confound removal.

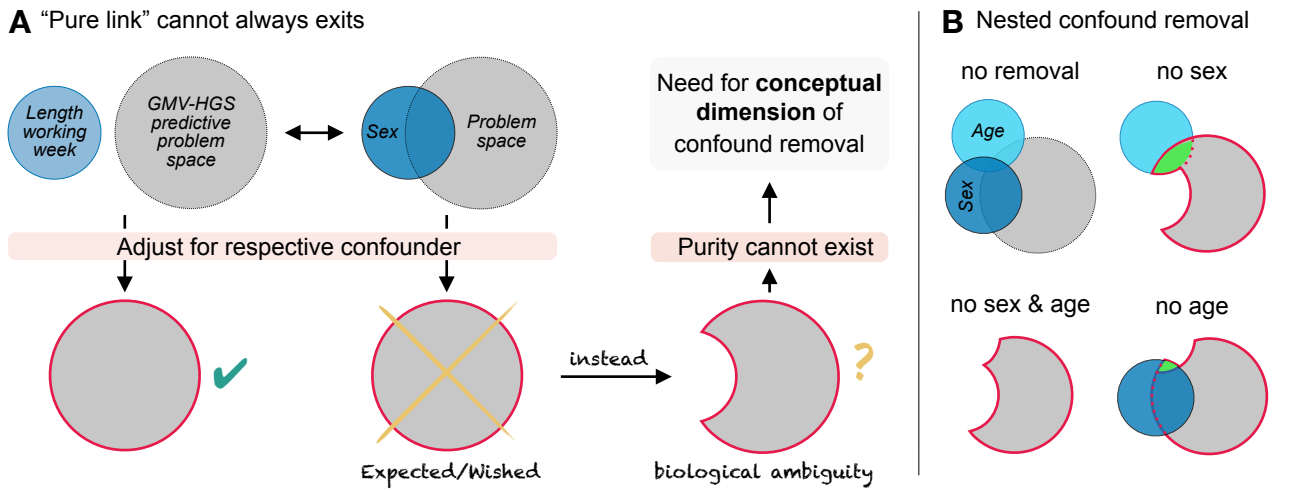

Figure 2. The concept of biological overlap of a potential confounder with the predictive problem space. **A)** The grey circle represents the set of biological attributes of the predictive problem space. The smaller blue circles visualize the set of biological attributes of the potential confounders "Length of working week" (brighter blue) and "sex" (darker blue). The red outlines surrounding the grey circle depict the "pure" biology of the problem space. In contrast to "length of working week", "sex" quite overlaps in its biological attributes with the GMV-HGS problem space. Therefore the wished "pure" biology of the problem space when removing sex as a confounder (red outline middle) cannot be reached. Instead the removal results in the peculiarly shaped red outline (bottom right). This new set of biological attributes of the GMV-HGS is biologically ambiguous and unclear in its interpretation. It motivates to acknowledge a conceptual dimension of confound removal to determine the reachable biological "purity" when adjusting for confounders. **B)** Nested overlap of confounders with the problem space and respective impact of removal. Sex and age demonstrate a nested overlap with the GMV-HGS problem-space. While in a non-nested scenario, removing one confounder alone would reveal the entire impact of removal, in the nested sex-age scenario illustrated here, only the joint removal reveals their full impact. Note: This figure employs Venn diagrams to visualize the degree of "shared biology" between variables and the consequences for model interpretation. This should not be confused with "shared variance" in a statistical sense.

3. Integrating the statistical and conceptual level of confound removal

3.1. The *Confound Continuum*

Beyond the extreme examples of "length of working week" and "sex", numerous further potential confounders exist for the GMV-HGS prediction task. Statistically, these can be ordered along a vertical axis based on increasing (absolute) strength of statistical association with the prediction task, as introduced in **Figure 1**.

Conceptually, further potential confounders can be ordered along a horizontal axis based on their increasing overlap with the biological attributes of the problem space (**Figure 3A**). On this continuum, "length of working week" exemplifies a low biological overlap or link, followed by the further potential confounder "bone density". The latter likely has differing driving factors than both GMV and HGS, yet the possibility of a biological link cannot be entirely ruled out. Advancing in the direction of increasing overlap, "systolic blood pressure" potentially shares driving factors with HGS, such as physical fitness, without a clearly identified pathway. "Sex" and hormonal composition almost reflect a 1:1 mapping of the same underlying biology, forming an example of a high biological link.

Integrating this horizontal conceptual axis (**Figure 3B**, blue) with the vertical statistical axis (**Figure 3B**, red) creates a two-dimensional (2D) orthogonal space (**Figure 3B**), emphasizing the independence of conceptual and statistical considerations. This independence becomes particularly evident for the off-diagonal variables in the 2D grid (**Figure 3B**, grey shaded areas). For instance, although "systolic blood pressure" only correlated marginally with GMV, biologically both may be influenced by a third factor such as physical fitness. Conversely, "length of working week" was correlated with HGS yet lacks evident overlap of biological attributes. The statistical dimension determines the amount of shared signal and thereby either ensures that no confounder information is inadvertently introduced to the data (no shared signal) or reveals which variables' removal may enhance the signal to noise ratio (high shared signal). While statistical evaluations are essential, they cannot address the semantic meaningfulness of removing confounders. Put differently, they cannot assess

the biomedical validity of confound-adjusted features, targets and resultant models and predictions. In contrast, the conceptual dimension offers valuable insights in the achievable purity of a feature-target (here: brain-behavioural) link, complementing statistical approaches with domain expertise and biological knowledge. Together, these dimensions dissect the different roles of potential confounding variables for a specific predictive problem from complementary viewpoints.

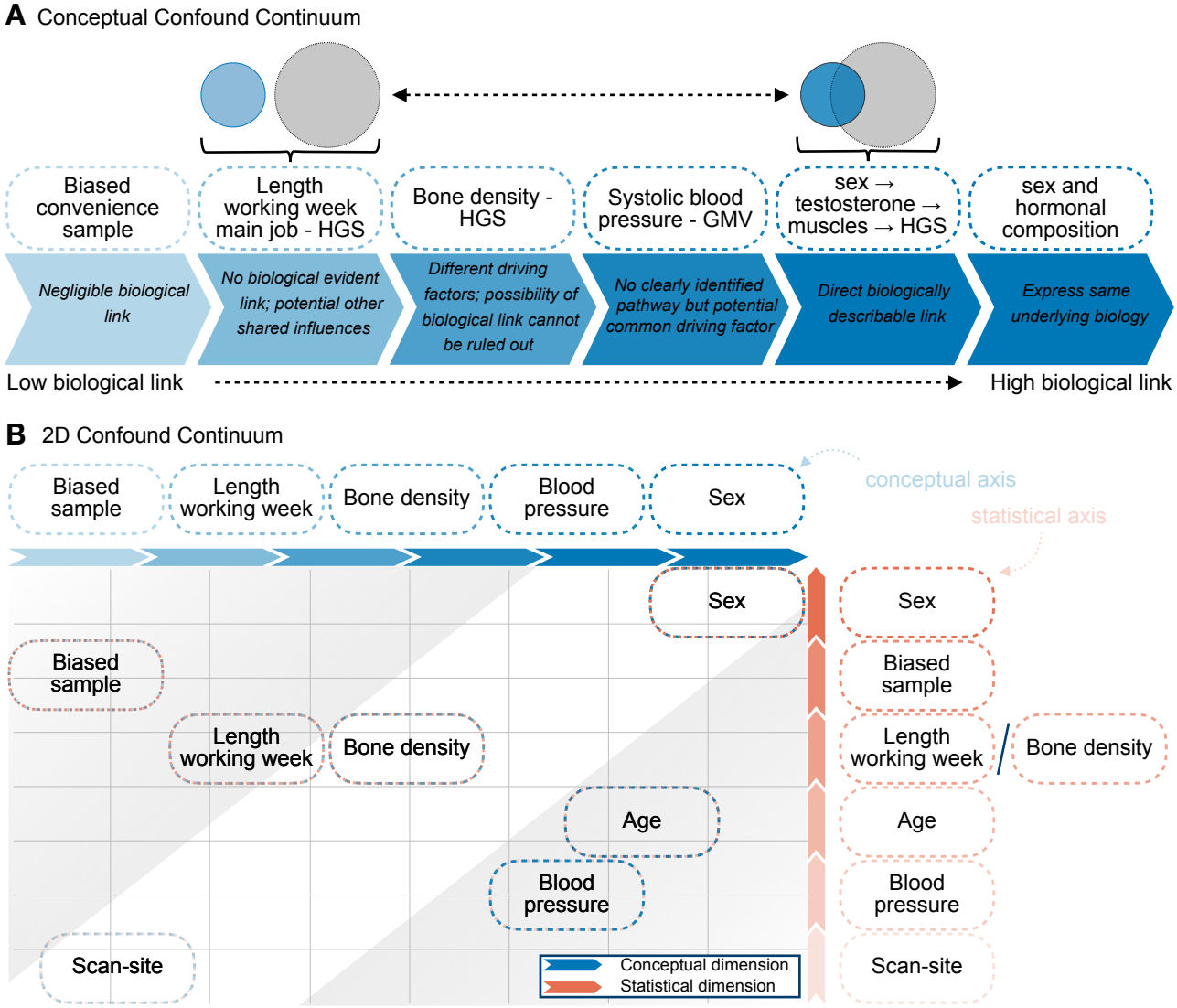

Figure 3. Conceptual and two-dimensional *Confound Continuum*. **A)** The conceptual *Confound Continuum* is the logical consequence of a gradual overlap in biological attributes between a set of potential confounders and the GMV-HGS predictive problem space. It is formed by ordering potential confounders in increasing order of shared biology with the problem space. The two example confounders "length of working week" and "sex" introduced before form two somewhat extremes on this conceptual continuum. **B)** The conceptual axis from A) can be combined with forementioned statistical evaluations (**Figure 1**) to form a two-dimensional grid. Importantly, the two dimensions are independent, which becomes particularly obvious by the variables in the grey shaded off-diagonals. The two-dimensional assessment helps to dissect the different roles of potential confounding variables for a particular predictive problem.

3.2. Nested and cascadic influences

Potential confounding variables may exhibit nested or cascadic overlaps with the problem space. For instance, sex and age demonstrate a nested overlap with the GMV-HGS problem-space (**Figure 2B**, top left, "no removal"). Adjusting for age preserves the shared area of sex, age and the problem space (**Figure 2B**, bottom right, "no age": green area) because it is encompassed by the sex-problem-space overlap. Consequently, only a small section of the red problem space outline (visually spoken) is missing in **Figure 2B** (bottom right, "no

age"), preserving most interpretability of the problem space. With sex adjustment, a comparable scenario emerges, but with a somewhat higher impact due to the larger overlap of sex-problem-space attributes (**Figure 2B**, top right, "no sex"). Adjusting for multiple confounders results in an additive removal effect in both nested and non-nested settings. However, in a non-nested scenario, removing one confounder alone would reveal the entire impact of removal. In contrast, in the nested sex-age example, only the joint removal reveals their full impact (**Figure 2B**, bottom left, "no sex & age").

Cascadic influences emerge because biomedical mechanisms usually form complex networks (see e.g.[57,58] for a formulation using directed acyclic graphs). For example, sex and hormones influence body fat composition. However, body fat composition in conjunction with sex can also influence hormones[59], which then can affect the biological cascade sex → testosterone → muscle growth → HGS. Body fat compositions may further overlap with respiratory performance, shaping additional factors such as physical fitness. Consequently, even seemingly unrelated variables may indirectly impact the actual relationship between GMV and HGS.

In summary, statistical and conceptual evaluations of confounder influences are independent but can be integrated as a two-dimensional grid – the *Confound Continuum*. This framework emphasizes that biomedical and statistical validity are distinct but complementary concepts to enhance our understanding of the role of confounders in a predictive task. The *Confound Continuum* can facilitate informed decisions on confound removal, acknowledging problem-specific nuances.

4. Confound removal can influence predictions more than feature or algorithm choice

To illustrate the importance of considering confounding variables in predictive workflows, we conducted the GMV-HGS prediction based on cortical[60], subcortical[61] and cerebellar[62] GMV features. The "vanilla" model, without removing confounders and using a linear support vector regression (SVR), yielded a Pearson correlation between true and predicted HGS of $R^2 = 0.39$ ($r = .63$, **Figure 4A**, left). We compared this "vanilla" model with models that linearly regressed out confounders prevalent in the field (scan-site, age, and sex)[12,44]. Additionally, we examined the combined effect of sex and age to illustrate a nested (additive) scenario. The scan-site adjusted model performed similarly to the vanilla model ($R^2 = 0.40$, $r = 0.64$). Adjusting GMV for sex substantially reduced performance ($R^2=.03$, $r = 0.20$), while age adjustment had no effect ($R^2 = 0.39$, $r = 0.63$). However, removing both sex and age resulted in a pronounced drop in performance ($R^2 = -0.0$, $r = 0.08$, **Figure 4A**, right), suggesting a nested additive scenario where regressing out sex revealed the signal contributions of age in GMV.

The choice of both, features and learning algorithm plays a crucial role in neuroimaging predictive modelling. Features should provide sufficient information about the target variable, and different learning algorithms can capture different aspects of the feature-target relationship (e.g. linear vs. non-linear relations). Therefore, in neuroimaging predictive workflows, often the features and learning algorithms are tweaked to explore if other neuroimaging derivatives carry a stronger signal about the target or other learning algorithms can detect the relationship better. In our example, using functional connectivity (FC) features instead of GMV, maintained comparable accuracy ($R^2 = 0.34$, $r = 0.58$, **Figure 4C**), while cortical thickness (CT) features less good ($R^2 = 0.13$, $r = 0.36$). Tweaking the learning algorithm or its fine-tuning had minimal impact (**Figure 4C**). Importantly, these influences were observed without confound removal. Thus, the lower performance of CT does not necessarily indicate it contains less information about HGS but could imply that CT carries less information about sex (and age) compared to GMV and FC.

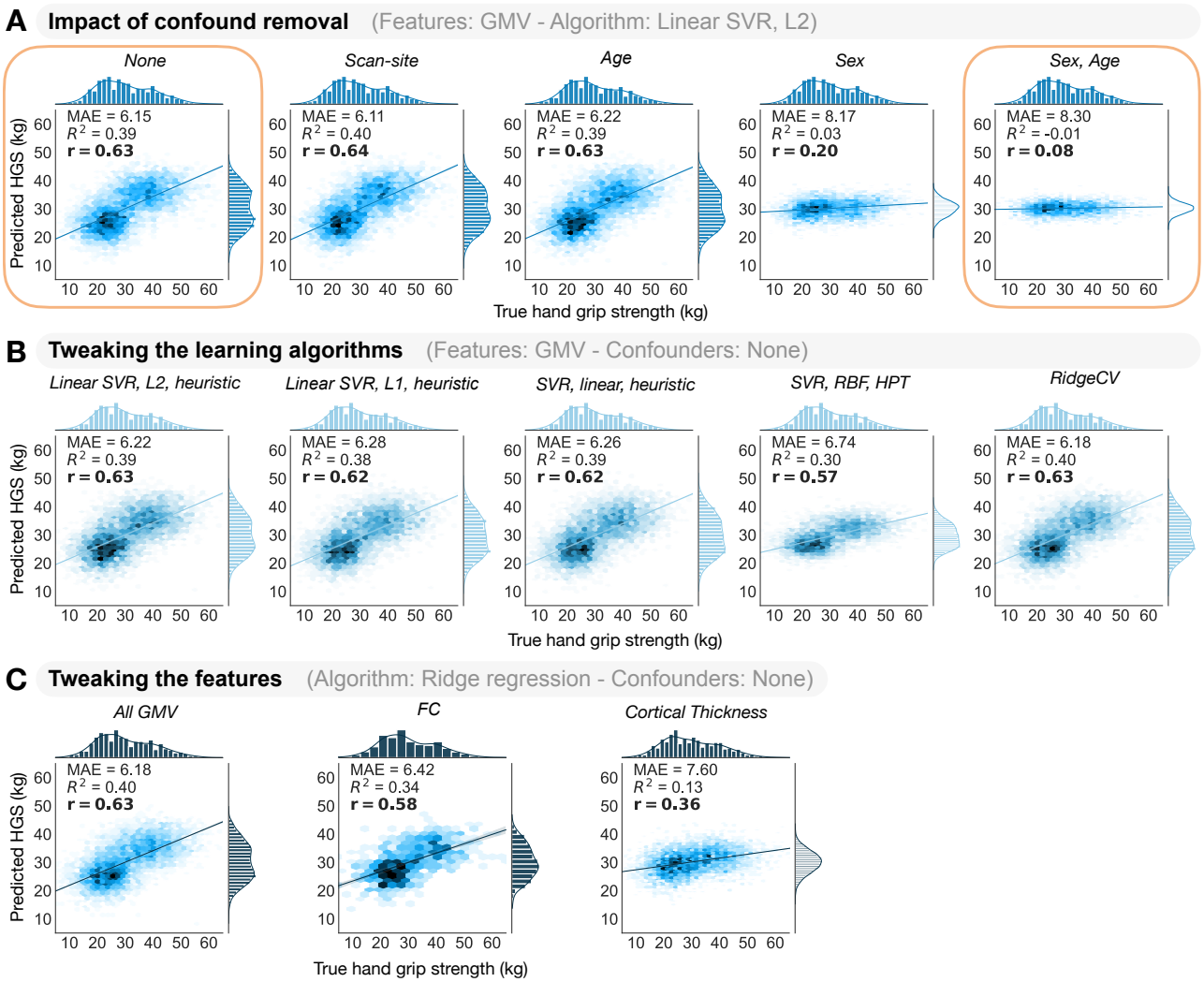

Figure 4. Impact of confound removal on a model's performance in (exemplary) comparison to tweaking neuroimaging features or learning algorithm. **A)** The impact of five different confound removal scenarios on the predictive performance ("vanilla", scan-site, age, sex, sex&age) with features and learning algorithm kept constant as GMV and Linear SVR with L2 loss and heuristic hyperparameter C, respectively. **B)** Five examples of differently (tuned) algorithms, while both feature choice (GMV) and confound removal scenario ("vanilla") were kept constant. **C)** Influence of feature choice (GMV, functional connectivity (FC), cortical thickness), with no confound removal ("vanilla") and hyperparameter tuned ridge regression. The orange boxes mark the two models differing most in performance.

To validate these findings, we additionally used confounders directly as features, with and without neuroimaging-derived features (**Figure 5**). Age and sex together as features (without "brain" features) outperformed models solely based on neuroimaging derived features (R^2 = 0.60, r = 0.77, **Figure 5C**, left). Adding GMV or CT to "sex & age" or "sex" as confound-features did not improve accuracy (**Figure 5B** & **C**). Incorporating FC alongside these two confound-feature setups even resulted in slight performance drops (R^2 = 0.37, r = 0.65 and R^2 = 0.36, r = 0.64, respectively, **Figure 5B** & **C**, right). In contrast, all brain-derived features contributed meaningfully to age as a confound-feature (**Figure 5A**). These insights align with **Figure 4**, emphasizing that the high performance of the HGS "vanilla" model is strongly driven by neural encodings of sex (and age) in the neuroimaging derived features.

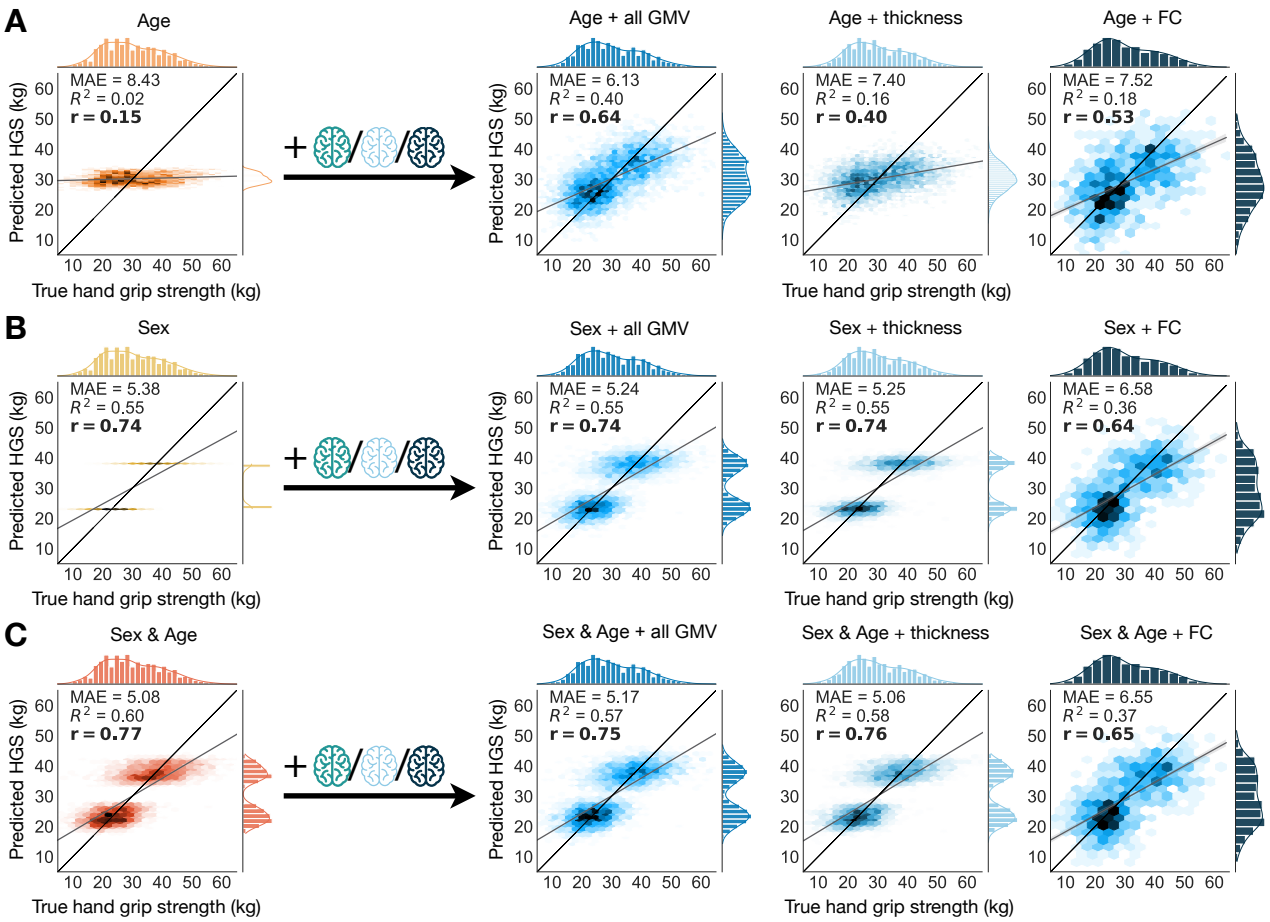

Figure 5. Confounders as features with and without brain features. **A)** For age as confound-feature (left), adding GMV, cortical thickness or FC, respectively adds information and increases the predictive performance for the target. **B)** and **C)** For Sex or Sex and Age as confound features, none of the additional brain features adds information about the target, which can be seen in same or even reduced predictive accuracy. Note: Although the distribution for the predicted HGS in **B** looks categorical, r = 0.74 shows the Pearson correlation coefficient, as both, the true and the predicted HGS are supposed to be a continuous outcome. The impression of a categorical distribution even when adding brain information shows even stronger that the prediction is driven by sex information.

While sex and age confounder adjustment significantly impacted predictive performance (r = .63 to r = .08), the most substantial difference due to feature or algorithm choice was only between GMV (r = .63) and CT (r = .36). This underscores that confounders can have a more pronounced impact on predictions than feature or learning algorithm selection. Selecting meaningful features and aligning algorithm choice with the assumed nature of the feature-target relationship is undoubtedly important. However, our results highlight that it is (at least) equally important to consider and understand the role of confounders in a predictive workflow.

5. Discussion

Precision medicine ML workflows are susceptible to context-dependent confounding influences. We differentiate between two overarching research endeavours, *high-performance* and *pure link*. Both require a nuanced understanding of confounders, either to avoid generalizability issues and identify potential covariate shifts (in *high-performance* case) or to determine the achievable purity of the problem space (in *pure link* case). We elaborated that such purity is difficult to achieve, if even reachable, in the case of biologically highly linked variables. To address the gradual nature of shared biology between potential confounders and a predictive problem space, we introduced a conceptual dimension of confound removal, ordering variables based on increasing biological link. This supplements statistical confounder evaluations by providing insights into biomedical implications of confound removal. The empirical HGS predictions underpinned the pivotal role of confounders in predictive workflows.

The substantial difference between the "vanilla" and the age-sex-adjusted model, raises the crucial question of which model is the correct one. Although this decision depends on the research endeavour (*high-performance* vs. *pure link*), yet the interpretation must align with this decision. The *high-performance* vanilla model predicts HGS decently but does not allow a statement about the finding of neural encodings of HGS. Additionally, also in this *high-performance* setup, training and test distributions must match to avoid covariate shift and enable transition from development into clinical practice, counteracting the AI chasm. Conversely, the sex-age-adjusted model may show lower performance but elucidates that sex and age encodings in GMV drive linear predictions of HGS. Despite lower accuracy, such models can enhance the understanding of (in this example) brain function beyond biologically overlapping other behavioural and phenotypical measures. Removing the influence of relevant variables, such as sex, uncovers smaller underlying signals and unmasks the necessity for deeper investigations. In fact, the nested additive effect of age in the GMV-HGS prediction would not have been discerned without removing the influence of sex.

The *Confound Continuum* aims to support informed confound removal decisions in a problem-dependent manner, bridging the gap between statistical and conceptual perspectives. It emphasizes that biomedical and statistical validity are distinct concepts and connects confound removal to model interpretation. In the realm of biology, no variable exists in complete isolation from others. Certain datasets might create the impression of some variables being biologically unrelated, but this likely reflects the inherent limitations of any dataset, which can only capture a finite number of measured variables. Therefore, it is crucial to dissect the interconnectedness of biological variables from a bio-conceptual perspective and combine this perspective with statistical data-insights to derive valid models and corresponding interpretations – a bridge provided by the *Confound Continuum*.

The necessity for integrating a bio-conceptual dimension with a statistical dimension of confound removal extends beyond neuroimaging predictive scenarios, being relevant for the entire domain of precision medicine. Despite successes in various prediction tasks, including cancer diagnosis and prognosis, inflammatory disease risk prediction, Alzheimer's disease progression prediction, identification of hyperkalaemia from electrocardiograms or identification of genetic conditions from facial appearance[63], the integration of AI in clinical practice still faces significant challenges. Most AI systems are far from achieving reliable generalizability, a prerequisite for clinical applicability[63]. For example, prognostic breast cancer models or predictive models for schizophrenia treatment outcomes only perform well in internal validation cohorts, but fail in external validation cohorts or trials[64,65], i.e. the models fail to generalize to unseen data. This is problematic because accuracy achieved during model development does not necessarily represent clinical efficacy, particularly if high performances were achieved by neglecting confounder influences. While various factors contribute to failure in generalizability, confounding influences, such as technical differences between sites, variations in local clinical practices or differing demographics between patients in different hospitals, represent a major obstacle. Undoubtedly, high performance is crucial for constructing useful clinical AI systems. Nevertheless, there will be always a degree of uncertainty and error in predictive models, so that it is essential to understand the strengths and limitations of AI tools[66]. Recognizing the impact of confounders on predictive models and particularly their biological and clinical meaning, as supported by the conceptual dimension of the *Confound Continuum*, can contribute to a more nuanced understanding and future development of these tools.

The present study has a limited statistical scope, focusing on correlations for the statistical *Confound Continuum* and linear regression for confounder adjustment. Exploring non-linear methods was not the intention as that can be found elsewhere (e.g.[18]). Although current non-statistical guidance for confound removal in brain-behavioural predictive modelling is limited, the conceptual considerations are not meant as a step-by-step guide to determine which confounders to remove. Future research in biomedicine or causal modelling may offer more specific guidance. Instead, it aims to raise awareness of the non-statistical biomedical dimension of confound removal, emphasizing the importance of appropriate model and results interpretation and of providing biomedical meaning and validity to predictive outcomes.

6. Conclusion

In data-driven predictive models, confounder decisions often rely solely on statistical and historical criteria. We here want to stress the necessity of supplementing statistical approaches with domain expertise and biomedical knowledge. The introduced 2D *Confound Continuum* integrates statistical and conceptual considerations, aiding in assessing the statistical and biomedical role of specific confounders for a particular research question and predictive context. When a statistical relationship exists between a confounder and the feature(s)/target, both removing or not removing potential confounders holds validity. However, the chosen strategy must match the intended goal of the model and interpretation of outcomes must differ accordingly. While reaching high performances is important, reflecting on the meaning of a model and how it can help to improve the medical field and our understanding of biomedical mechanisms is at least as important. The *Confound Continuum* fosters such an overall perspective, supporting accurate model interpretation and discouraging uncritical model employment.

7. Acknowledgments

This research has been conducted using data from UK Biobank, a major biomedical database (www.ukbiobank.ac.uk). This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 431549029 - Collaborative Research Centre CRC1451 on motor performance project B05.

8. References

1. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *Npj Digit Med*. 2021;4(1):153. doi:10.1038/s41746-021-00521-5

2. Darcy AM, Louie AK, Roberts LW. Machine Learning and the Profession of Medicine. *JAMA*. 2016;315(6):551. doi:10.1001/jama.2015.18421

3. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243. doi:10.1136/svn-2017-000101

4. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov*. 2021;11(4):900-915. doi:10.1158/2159-8290.CD-21-0090

5. Subramanian M, Wojtusciszyn A, Favre L, et al. Precision medicine in the era of artificial intelligence: implications in chronic disease management. *J Transl Med*. 2020;18(1):472. doi:10.1186/s12967-020-02658-5

6. Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(3):223-230. doi:10.1016/j.bpsc.2017.11.007

7. Citerio G. Big Data and Artificial Intelligence for Precision Medicine in the Neuro-ICU: Bla, Bla, Bla. *Neurocrit Care*. 2022;37(S2):163-165. doi:10.1007/s12028-021-01427-6

8. Heinrichs B, Eickhoff SB. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum Brain Mapp*. 2020;41(6):1435-1444. doi:10.1002/hbm.24886

9. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *Npj Digit Med*. 2018;1(1):40, s41746-018-0048-y. doi:10.1038/s41746-018-0048-y

10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

11. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*. 2017;145:137-165. doi:10.1016/j.neuroimage.2016.02.079

12. Benkarim O, Paquola C, Park B yong, et al. The Cost of Untracked Diversity in Brain-Imaging Prediction. *bioRxiv*. Published online June 2021:34. doi:https://doi.org/10.1101/2021.06.16.448764

13. Pulini AA, Kerr WT, Loo SK, Lenartowicz A. Classification Accuracy of Neuroimaging Biomarkers in Attention-Deficit/Hyperactivity Disorder: Effects of Sample Size and Circular Analysis. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2019;4(2):108-120. doi:10.1016/j.bpsc.2018.06.003

14. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci*. 2017;20(3):365-377. doi:10.1038/nn.4478

15. Kapoor S, Narayanan A. Leakage and the Reproducibility Crisis in ML-based Science. Published online July 14, 2022. Accessed January 31, 2023. http://arxiv.org/abs/2207.07048

16. Organization WH, others. Ethics and governance of artificial intelligence for health: WHO guidance. Published online 2021.

17. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci*. 2016;19(3):404-413. doi:10.1038/nn.4238

18. Alfaro-Almagro F, McCarthy P, Afyouni S, et al. Confound modelling in UK Biobank brain imaging☆. Published online 2021:17.

19. Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF. *Controlling for Effects of Confounding Variables on Machine Learning Predictions*. Bioinformatics; 2020. doi:10.1101/2020.08.17.255034

20. Weinberger DR, Radulescu E. Finding the Elusive Psychiatric "Lesion" With 21st-Century Neuroanatomy: A Note of Caution. *Am J Psychiatry*. 2016;173(1):27-33. doi:10.1176/appi.ajp.2015.15060753

21. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*. 2012;5(2):79-83.

22. Chyzhyk D, Varoquaux G, Milham M, Thirion B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*. 2022;11:giac014. doi:10.1093/gigascience/giac014

23. Jager KJ, Zoccali C, MacLeod A, Dekker FW. Confounding: What it is and how to deal with it. *Kidney Int*. 2008;73(3):256-260. doi:10.1038/sj.ki.5002650

24. Kostro D, Abdulkadir A, Durr A, et al. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *NeuroImage*. 2014;98:405-415. doi:10.1016/j.neuroimage.2014.04.057

25. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the Mediation, Confounding and Suppression Effect. *Prev Sci*. 2000;1(4):9.

26. McNamee R. Confounding and confounders. *Occup Environ Med*. 2003;60(3):227-234. doi:10.1136/oem.60.3.227

27. Rao A, Monteiro JM, Mourao-Miranda J. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*. 2017;150:23-49. doi:10.1016/j.neuroimage.2017.01.066

28. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. *Nat Commun*. 2020;11(1):6010. doi:10.1038/s41467-020-19784-9

29. Geerligs L, Tsvetanov KA, Cam-CAN, Henson RN. Challenges in measuring individual differences in functional connectivity using fMRI: The case of healthy aging: Measuring Individual Differences Using fMRI. *Hum Brain Mapp*. 2017;38(8):4125-4156. doi:10.1002/hbm.23653

30. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*. 2012;59(3):2142-2154. doi:10.1016/j.neuroimage.2011.10.018

31. Satterthwaite TD, Wolf DH, Loughead J, et al. Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage*. 2012;60(1):623-632. doi:10.1016/j.neuroimage.2011.12.063

32. Spisak T. Statistical quantification of confounding bias in predictive modelling. Published online November 1, 2021. Accessed January 31, 2023. http://arxiv.org/abs/2111.00814

33. Bugg JM, Zook NA, DeLosh EL, Davalos DB, Davis HP. Age differences in fluid intelligence: Contributions of general slowing and frontal decline. *Brain Cogn*. 2006;62(1):9-16. doi:10.1016/j.bandc.2006.02.006

34. Hartshorne JK, Germine LT. When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychol Sci*. 2015;26(4):433-443. doi:10.1177/0956797614567339

35. Horn (1967) - age differences in fluid and crystallized intelligence.pdf.

36. Kahlert J, Gribsholt SB, Gammelager H, Dekkers OM, Luta G. Control of confounding in the analysis phase – an overview for clinicians. *Clin Epidemiol*. 2017;Volume 9:195-204. doi:10.2147/CLEP.S129886

37. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown; 2017.

38. Abdulkadir A, Ronneberger O, Tabrizi SJ, Klöppel S. Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. In: *2014 International Workshop on Pattern Recognition in Neuroimaging*. IEEE; 2014:1-4.

39. Dukart J, Schroeter ML, Mueller K, The Alzheimer's Disease Neuroimaging Initiative. Age Correction in Dementia – Matching to a Healthy Brain. Valdes-Sosa PA, ed. *PLoS ONE*. 2011;6(7):e22193. doi:10.1371/journal.pone.0022193

40. Snoek L, Miletić S, Scholte HS. How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*. 2019;184:741-760. doi:10.1016/j.neuroimage.2018.09.074

41. Rao A, Monteiro JM, Ashburner J, et al. A comparison of strategies for incorporating nuisance variables into predictive neuroimaging models. In: *2015 International Workshop on Pattern Recognition in Neuroimaging*. ; 2015:61-64.

42. Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523-1536. doi:10.1038/nn.4393

43. Stein JL, Medland SE, Vasquez AA, et al. Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet*. 2012;44(5):552-561.

44. Weinstein SM, Davatzikos C, Doshi J, Linn KA, Shinohara RT, For the Alzheimer's Disease Neuroimaging Initiative. Penalized decomposition using residuals (PeDecURe) for feature extraction in the presence of nuisance variables. *Biostatistics*. Published online August 11, 2022:kxac031. doi:10.1093/biostatistics/kxac031

45. Westfall J, Yarkoni T. Statistically Controlling for Confounding Constructs Is Harder than You Think. Tran US, ed. *PLOS ONE*. 2016;11(3):e0152719. doi:10.1371/journal.pone.0152719

46. Wachinger C, Rieckmann A, Pölsterl S. Detect and correct bias in multi-site neuroimaging datasets. *Med Image Anal*. 2021;67:101879. doi:10.1016/j.media.2020.101879

47. Smith SM, Nichols TE. Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*. 2018;97(2):263-268. doi:10.1016/j.neuron.2017.12.018

48. Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biom Bull*. 1946;2(3):47. doi:10.2307/3002000

49. Hamdan S, Love BC, von Polier GG, et al. Confound-leakage: confound removal in machine learning leads to leakage. *GigaScience*. 2023;12.

50. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

51. Bobos P, Nazari G, Lu Z, MacDermid JC. Measurement Properties of the Hand Grip Strength Assessment: A Systematic Review With Meta-analysis. *Arch Phys Med Rehabil*. 2020;101(3):553-565. doi:10.1016/j.apmr.2019.10.183

52. Gell M, Eickhoff SB, Omidvarnia A, et al. The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions. *bioRxiv*. Published online February 10, 2023. doi:10.1101/2023.02.09.527898

53. Alonso AC, Ribeiro SM, Luna NMS, et al. Association between handgrip strength, balance, and knee flexion/extension strength in older adults. Sergi G, ed. *PLOS ONE*. 2018;13(6):e0198185. doi:10.1371/journal.pone.0198185

54. Hamdan S, Love BC, von Polier GG, et al. Confound-leakage: Confound removal in machine learning leads to leakage. *ArXiv Prepr ArXiv221009232*. Published online 2022.

55. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. *Dataset Shift in Machine Learning*. Mit Press; 2008.

56. Huyen C. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. First edition. O'Reilly Media, Inc; 2022.

57. Wysocki AC, Lawson KM, Rhemtulla M. Statistical Control Requires Causal Justification.

58. Chen G, Cai Z, Taylor PA. Through the lens of causal inference: Decisions and pitfalls of covariate selection. Published online January 12, 2024. doi:10.1101/2024.01.11.575211

59. Tchernof A, Després JP, Bélanger A, et al. Reduced testosterone and adrenal C19 steroid levels in obese men. *Metabolism*. 1995;44(4):513-519. doi:10.1016/0026-0495(95)90060-8

60. Schaefer A, Kong R, Gordon EM, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex*. 2018;28(9):3095-3114. doi:10.1093/cercor/bhx179

61. Tian Y, Margulies DS, Breakspear M, Zalesky A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat Neurosci*. 2020;23(11):1421-1432. doi:10.1038/s41593-020-00711-6

62. Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N. A probabilistic MR atlas of the human cerebellum. *NeuroImage*. 2009;46(1):39-46. doi:10.1016/j.neuroimage.2009.01.045

63. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. doi:10.1186/s12916-019-1426-2

64. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. *BMC Cancer*. 2019;19(1):230. doi:10.1186/s12885-019-5442-6

65. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science*. 2024;383(6679):164-167. doi:10.1126/science.adg8538

66. Cascella M, Montomoli J, Bellini V, et al. Crossing the AI Chasm in Neurocritical Care. *Computers*. 2023;12(4):83. doi:10.3390/computers12040083