Statistical learning for complex data to enable precision medicine strategies

Zwep, L.B.

Chapter 9

Nederlandse Samenvatting

Samenvatting

Om de behandeling met geneesmiddelen van individuele patiënten te verbeteren is het cruciaal om de variabiliteit tussen patiënten en hun reactie op verschillende geneesmiddelen en doseringsschema's, beter te begrijpen. In dit proefschrift hebben wij onderzoek gedaan naar verschillende factoren en statistische methodes om dergelijke variabiliteit tussen patiënten kunnen voorspellen. In de afgelopen jaren zijn er grote ontwikkelingen geweest op het gebied van meettechnologie waarmee zeer efficiënt genetische of biochemische profielen kunnen worden bepaald in patiëntmateriaal, die mogelijk als zogenaamde biomarker bij kunnen dragen aan het voorspellen van variatie tussen patiënten. Dergelijke profielen zijn hoog-dimensionaal: voor elk patiëntmonster kunnen er vele honderden tot duizenden unieke genetische of biochemische waarden bepaald worden. Bovendien worden deze hoog-dimensionale gegevens ook steeds meer longitudinaal (over de tijd) gemeten in patiënten om zo verdere inzichten in verloop van ziekte en therapieresponse te kunnen verkrijgen. Echter, het afleiden van bruikbare inzichten uit dergelijke gegevens is vooralsnog een uitdaging (Sectie 1). Om tot deze inzichten te komen, zijn statische methoden voor longitudinale en hoog-dimensionale gegevens nodig. Met deze methoden kan gezocht worden naar biomarkers die de variabiliteit tussen patiënten verklaren, in termen van reactie op de behandeling en klinisch verloop van de ziekte (Sectie 2). Daarnaast zijn deze methoden nodig om klinische vragen te beantwoorden met behulp van gegevens uit de medische praktijk, die in de reguliere gezondheidszorg en onderzoeksinstituten verkregen worden (Sectie 3).

9.1 Verschillende gegevens in farmaceutisch onderzoek

Het doel van klinische studies is om kennis te vergaren over de response van patiënten op een behandeling. Echter, niet alle soorten patiënten, medicatie en bijwerkingen worden door middel van klinische studies onderzocht, dit vanwege de uitdagingen en kosten die komen kijken, bij bijvoorbeeld kleine populaties of weinig voorkomende effecten. De beslissingen over dit soort behandelingen en patiëntengroepen kunnen in dat geval genomen worden met behulp van gegevens uit de gezondheidszorg en metingen die thuis kunnen worden gedaan (Morrato et al., 2007; Swift et al., 2018). In Hoofdstuk 2, beschreven we verschillende gegevenstypen die gebruikt kunnen worden voor het maken van klinische beslissingen. Ook beschreven we welke mogelijkheden en uitdagingen deze gegevenstypen met zich meebrengen. Hierbij ligt de focus op geneesmiddeleffecten in kinderen, omdat klinische studies daar vaak belemmerd worden door problemen met het rekruteren van patiënten (Brussee et al., 2016).

Gegevens uit de medische praktijk kunnen niet op dezelfde manier geanalyseerd worden als gegevens uit klinische studies. Dat zorgt ervoor dat er voor het analyseren van deze gegevens aangepaste of uitgebreide statistische methoden, zoals

machine learning methoden, nodig zijn. Het gebruik van zulke methoden levert een extra uitdaging op, omdat de uitkomsten van dit soort methoden vaak moeilijk te interpreteren zijn, en het bij het maken van medische beslissingen sterk van belang is om uit te kunnen leggen waarom deze beslissingen worden genomen (Knoppers & Thorogood, 2017).

9.2 Zoektocht naar biomarkers

De vooruitgang in meettechnieken voor het bepalen van moleculaire waarden maakt het mogelijk om de moleculaire samenstelling van een mens preciezer te meten, bijvoorbeeld uit bloed- of urinemonsters. Het is mogelijk om gegevens over moleculen, zoals metabolieten, RNA en DNA, te meten om een patiënt met grote precisie te beschrijven (Pearson, 2016). Deze zogeheten 'omics' gegevens maken het mogelijk om veel meer factoren te bestuderen dan voorheen, maar dat maakt de analyse voor het detecteren van voorspellende biomarkers ook een stuk complexer (Depledge et al., 1993), zeker in het geval van tijdgerelateerde veranderingen. De statistische technieken die op dit moment beschikbaar zijn, zijn vaak niet toepasbaar op gegevens die zowel hoog-dimensionaal als longitudinaal zijn. De meeste methoden zijn ontworpen om maar met één van deze twee aspecten om te kunnen gaan.

In Hoofdstuk 3, identificeerden we biomarkers voor het volgen van de reactie op verschillende kankerbehandelingen uit hoog-dimensionale DNA gegevens, die gemeten waren in muismodellen met kankercellen getransplanteerd uit patiënten. De gebruikte gegevens komen uit een gepubliceerde studie (Gao et al., 2015), waarbij DNA eigenschappen van kankercellen werden bepaald en vervolgens de tumorgroei over de tijd werd gemeten in muizen die behandeld waren met verschillende geneesmiddelen. In onze analyse hebben wij als eerste stap een longitudinaal tumorgroei inhibitie model ontwikkeld om de tumorgroei te beschrijven met behulp van drie karakteristieken: de groeisnelheid van de tumor (k_g), de sterkte van het geneesmiddeleffect (k_d) en de ontwikkeling van resistentie tegen het geneesmiddel (k_r). Met het model worden zowel populatie parameters als de individuele parameters (empirische Bayes schatters) bepaald. In de tweede stap werden de schattingen van deze individuele parameters als uitkomst in een zogenaamde lasso regressie gebruikt en werden ze gerelateerd aan de hoog-dimensionale DNA gegevens. We vonden door middel van een kruisvalidatie dat een deel van de variantie in geneesmiddelresponse verklaard kon worden met behulp van de DNA gegevens. Het gebruik van een biologisch geïnformeerde groep-lasso maakte het mogelijk om DNA effecten op het niveau van de biologisch processen te onderzoeken.

Het modelleren van hoog-dimensionale gegevens heeft als risico dat er zogenaamde overfitting kan optreden, waarbij de hoeveelheid variabelen ervoor kan zorgen dat de correlerende effecten niet geschat kunnen worden tussen de optredende ruis. Dit probleem kan ook voorkomen in de context van farmacometrische modellen, dus ook bij het schatten van parameters uit hoog-dimensionale DNA gegevens. Het voorspellen van de tumorgroei in de twee beschreven stappen omzeilt deze beperk-

ing. Een mogelijk risico van de door ons ontwikkelde twee-stapmethode voor biomarkeridentificatie is dat fouten in modelvoorspelling in de eerste stap kunnen worden uitvergroot in de tweede stap. Om het risico hierop te verkleinen is het dus van belang om de fout in beiden stappen aandachtig te bestuderen.

In Hoofdstuk 4 hebben we potentiële biomarkers voor het klinisch verloop, de combinatie tussen de reactie op de behandeling en het ziekteverloop, onderzocht in longitudinale metabolomische gegevens van in het ziekenhuis opgenomen patiënten met een thuis opgelopen longontsteking (community acquired pneumonia, CAP). We gebruikten de dimensie reductie methode principale component analyse om patronen van metabolieten over de tijd te verkennen en de relatie tussen de biochemische klassen van de metabolieten en het klinisch verloop te bestuderen. We berekenden de correlaties tussen de metabolieten en twee maten voor het klinisch verloop: de CURB score, een score die aangeeft hoe ziek de patiënt is bij aankomst in het ziekenhuis, en het aantal dagen van het verblijf in het ziekenhuis, welke aangeeft hoelang het duurde voor een patiënt om te herstellen. De patronen van de metabolieten veranderden duidelijk over de tijd, wat aangeeft hoe belangrijk het is om longitudinale metabolomische gegevens te bestuderen in patiënten met CAP. We identificeerden verschillende biochemische metabolietklassen die gerelateerd zijn aan het klinisch ziekteverloop, zoals de triglyceriden en de lyso-fosfatidylcholine klassen.

9.3 Gegevens uit de medische praktijk

In de gezondheidszorg worden routinematig zorggegevens van patiënten verzameld, om de hun gezondheid en hun reactie op de behandeling te volgen. Deze routinematig verzamelde zorggegevens worden tegenwoordig voor langere tijd opgeslagen, en vervolgens vaak beschikbaar gemaakt voor onderzoek. Dit soort gegevens, uit de medische praktijk, kunnen bijdragen aan het begrip van verschillende onderwerpen in de farmaceutische wetenschap.

Antibiotica-resistentie vormt een bedreiging voor de gezondheid van mensen wereldwijd. De mate van de dreiging en de ontwikkeling hiervan worden gemonitord door het bepalen van de mate van antibiotica-resistentie in patiënten opgenomen in ziekenhuizen (van der Kuil et al., 2017). Voor het verzamelen van deze gegevens wordt meestal de minimaal inhiberende concentraties (MIC) van de uit patiënten geïsoleerde bacteriële stammen bepaald en opgeslagen. Collaterale sensitiviteit (CS) is een proces waarbij in een populatie van pathogenen, de blootstelling aan één medicijn de resistentie tegen een tweede medicijn vermindert. Dit mechanisme kan bruikbaar kan zijn voor het bestrijden van resistente pathogenen (Aulin et al., 2021). De gemeten MIC waarden in de populatie kunnen gebruikt worden om deze CS te detecteren. Dit fenomeen is al meermaals gedetecteerd in proeven in een laboratoriumomgeving, maar de kennis over het voorkomen van CS in de klinische context is zeer beperkt.

In Hoofdstuk 5 hebben we een methode voorgesteld om collaterale effecten te kwantificeren in MIC gegevens (Zwep et al., 2021). Deze methode kwantificeert de

effecten als log_2 fold change, wat een interpreteerbare maat is die ook in laboratoriumproeven gebruikt wordt. Dit maakt het mogelijk om een directe vergelijking tussen laboratoriumproeven en de klinische praktijk te maken in de grootte van het effect. Deze maat werd gebruikt in Hoofdstuk 6 voor het kwantificeren van CS in grootschalige MIC bewakingsgegevens uit verschillende gegevensbronnen. Hieruit kwam naar voren dat CS leek voor te komen in de klinische praktijk, maar het was moeilijk om specifieke patronen over de verschillende soorten en antibiotica te detecteren. Dit gebrek aan generaliseerbaarheid maakt het lastig om CS in de klinische praktijk toe te passen (Nichol et al., 2019).

Routinematig verzamelde zorggegevens van patiënten bevatten verschillende typen covariaten, zoals biomarker concentraties, demografische gegevens en andere karakteristieken van de patiënten (Currie & MacDonald, 2000). In farmacometrische modellen worden deze karakteristieken vaak gebruikt om variabiliteit tussen patiënten te verklaren en voorspellen, wat mogelijkheden biedt tot het uitbreiden van de voorspellingen naar speciale patiëntengroepen. Farmacometrische simulaties vereisen het simuleren van deze covariaten, maar het delen van deze gevoelige patiëntengegevens tussen zorgverleners en onderzoeksgroepen is vaak niet mogelijk, vanwege het beschermen van de privacy van de patiënten.

In Hoofdstuk 7 stellen wij het gebruik van zogenaamde copulas voor als een geschikte methode voor het simuleren van virtuele patiënten. Copulas zijn multivariate verdelingsfuncties die gedeelde verdelingen kunnen beschrijven en bieden daarmee een flexibele manier om covariaten sets van patiënten te beschrijven en te simuleren. De gedeelde verdelingsfunctie van meerdere covariaten beschrijft de afhankelijkheid tussen de verschillende covariaten. De meeste covariaten zijn niet onafhankelijk van elkaar, dus het vatten van deze afhankelijkheid is cruciaal voor het simuleren van realistische virtuele patiënten. Onze studie liet zien dat we met copulas in staat zijn om realistische patiëntengroepen te simuleren (Zwep et al., 2022). Het is daarmee mogelijk om verschillende patiëntengroepen te simuleren, wat het extrapoleren van de resultaten uit modellen naar specifieke patiëntengroepen mogelijk te maakt.

De copulas worden geschat aan de hand van gegevens die niet altijd beschikbaar zijn voor onderzoekers, maar als de copulas geschat worden bij de onderzoeksgroep die de gegevens in bezit heeft kunnen de copulas gedeeld worden met andere onderzoekers, zodat zij de populatie kunnen simuleren en bestuderen zonder de oorspronkelijke gegevens in bezit te krijgen. Op deze manier kunnen covariaten van patiënten gedeeld worden tussen onderzoeksgroepen, zonder de privacy van de patiënt te schaden.

9.4 Conclusies

Om geneesmiddelgebruik te personaliseren is het nodig om de variabiliteit in de reacties van patiënten op een behandeling te ontrafelen. Hiervoor is het gebruik van statistische methoden die het gebruik van verschillende soorten gegevens, zoals

hoog-dimensionale omics, maar ook routinematig verzamelde zorggegevens, van essentieel belang. Integratie van state-of-the-art statistische methoden met farmacometrische modellen faciliteert farmacologisch onderzoek, door het mogelijk maken van gebruik van meerdere soorten gegevens, waarbij de interpreteerbaarheid van de farmacologische modellen behouden blijft.

Referenties

Aulin, L. B. S., Liakopoulos, A., van der Graaf, P. H., Rozen, D. E., & van Hasselt, J. G. C. (2021, Sep). Design principles of collateral sensitivity-based dosing strategies. *Nature Communications*, *12*(1). Retrieved from https://doi.org/10.1038%2Fs41467-021-25927-3 doi: 10.1038/s41467-021-25927-3

Brussee, J. M., Calvier, E. A. M., Krekels, E. H. J., Välitalo, P. A. J., Tibboel, D., Allegaert, K., & Knibbe, C. A. J. (2016, Jun). Children in clinical trials: towards evidence-based pediatric pharmacotherapy using pharmacokinetic-pharmacodynamic modeling. *Expert Review of Clinical Pharmacology*, *9*(9), 1235–1244. Retrieved from https://doi.org/10.1080%2F17512433.2016.1198256 doi: 10.1080/17512433.2016.1198256

Currie, C. J., & MacDonald, T. M. (2000). Use of routine healthcare data in safe and cost-effective drug use. *Drug Safety*, *22*(2), 97–102. Retrieved from https://doi.org/10.2165%2F00002018-200022020-00002 doi: 10.2165/00002018-200022020-00002

Depledge, M. H., Amaral-Mendes, J. J., Daniel, B., Halbrook, R. S., Kloepper-Sams, P., Moore, M. N., & Peakall, D. B. (1993). The conceptual basis of the biomarker approach. In *Biomarkers* (pp. 15–29). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007%2F978-3-642-84631-1_2 doi: 10.1007/978-3-642-84631-1_2

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., ... Sellers, W. R. (2015, Oct). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine*, *21*(11), 1318–1325. Retrieved from https://doi.org/10.1038%2Fnm.3954 doi: 10.1038/nm.3954

Knoppers, B. M., & Thorogood, A. M. (2017, Aug). Ethics and big data in health. *Current Opinion in Systems Biology*, *4*, 53–57. Retrieved from https://doi.org/10.1016%2Fj.coisb.2017.07.001 doi: 10.1016/j.coisb.2017.07.001

Morrato, E. H., Elias, M., & Gericke, C. A. (2007, Dec). Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in australia, the UK and the USA. *Journal of Public Health*, *29*(4), 463–471. Retrieved from https://doi.org/10.1093%2Fpubmed%2Ffdm065 doi: 10.1093/pubmed/fdm065

Nichol, D., Rutter, J., Bryant, C., Hujer, A. M., Lek, S., Adams, M. D., ... Scott, J. G. (2019, Jan). Antibiotic collateral sensitivity is contingent on the repeatability of evolution. *Nature Communications*, *10*(1). Retrieved from https://doi.org/10.1038%2Fs41467-018-08098-6 doi: 10.1038/s41467-018-08098-6

Pearson, E. R. (2016, May). Personalized medicine in diabetes: the role of 'omics' and biomarkers. *Diabetic Medicine*, *33*(6), 712–717. Retrieved from https://doi.org/10.1111%2Fdme.13075 doi: 10.1111/dme.13075

Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., & Jadhav, P. R. (2018, May). Innovation at the intersection of clinical trials and real-world data science to advance patient care. *Clinical and Translational Science*, *11*(5), 450–460. Retrieved from https://doi.org/10.1111%2Fcts.12559 doi: 10.1111/cts.12559

van der Kuil, W. A., Schoffelen, A. F., de Greeff, S. C., Thijsen, S. F., Alblas, H. J., Notermans, D. W., ... and, T. L. (2017, Nov). National laboratory-based surveillance system for antimicrobial resistance: a successful tool to support the control of antimicrobial resistance in the netherlands. *Eurosurveillance*, *22*(46). Retrieved from https://doi.org/10.2807%2F1560-7917.es.2017.22.46.17-00062 doi: 10.2807/1560-7917.es.2017.22.46.17-00062

Zwep, L. B., Guo, T., Nagler, T., Knibbe, C. A., Meulman, J. J., & van Hasselt, J. C. (2022). Virtual patient simulation using copula modeling. *[in preparation]*.

Zwep, L. B., Haakman, Y., Duisters, K. L. W., Meulman, J. J., Liakopoulos, A., & van Hasselt, J. G. C. (2021, Sep). Identification of antibiotic collateral sensitivity and resistance interactions in population surveillance data. *JAC-Antimicrobial Resistance*, *3*(4). Retrieved from https://doi.org/10.1093%2Fjacamr%2Fdlab175 doi: 10.1093/jacamr/dlab175