Den kostnadsintensiva patientgruppen i VGR

Ett samarbetsprojekt mellan Regional Vårdanalys på Koncernstab Digitalisering & Kompetenscentrum AI på Sahlgrenska Universitetssjukhuset

Anna Rosén, Juulia Suvilehto, Lisa Sjöblom, Tove Mathiasson, Viktoria Karlsson

Mars 2024 – preliminär rapport

1 Sammanfattning

Kompetenscentrum AI (Sahlgrenska Universitetssjukhuset) och enhet regional vårdanalys (Koncernstab Digitalisering) vid Västra Götalandsregionen (VGR) genomförde tillsammans ett projekt med syfte att kartlägga fördelningen av VGRs kostnader för specialiserad vård. Genom att beräkna regionens sammanlagda vårdkostnader för enskilda patienter (baserat på KPP-data, kostnad per patient) identifierades för varje år mellan 2015 och 2022 en grupp med så kallade kostnadsintensiva patienter, här definierat som de 5 % av patienterna som medfört högst kostnader för sjukhusen under året. En tydlig trend visade att gruppen av kostnadsintensiva patienter konsekvent konsumerar ungefär 50 % av sjukhusens totala budget och en deskriptiv analys visade på vissa utstickande egenskaper hos gruppens patienter. Ett försök gjordes sedan att ytterligare karaktärisera de kostnadsintensiva patienterna med hjälp av klusteranalys (HDBSCAN). Försöket resulterade i 20 kluster där vissa kliniska profiler kunde urskiljas. Förhoppningen är att analysen ska kunna ligga till grund för framtida arbete som syftar till att minimera onödiga utgifter kopplade till vissa patientgrupper och därmed leda till en effektivare användning av vårdens resurser samt en ökad vårdkvalitet.

1.1 Tack

Tack till Per Sjöli, regionutvecklare vid regional vårdanalys, för hjälp med datasammanställning och expertkunskaper kring KPP-data. Tack även till allmänläkare Emil Johansson vid regional vårdanalys för tankar kring rapporten och till Mikko Seppänen, avdelningsöverläkare på enheten för sällsynta sjukdomar (Nya barnsjukhuset, Helsingfors universitetssjukhus) som bidrog till tolkningen av kluster genom att dela med sig av sina medicinska kunskaper. Tack till region Västmanland för utbyte av lärdomar och till Vinnova som genom sin finansiering möjliggjorde genomförandet.

1.2 Begreppslista för maskininlärning

Datapunkt - All information som är samlad om en instans av något som ska undersökas, t.ex. en individ. Representeras ofta som en rad i en tabell.

Dataset - En strukturerad samling av datapunkter som delar variabler. Representeras ofta som en tabell där varje rad motsvarar en unik datapunkt och varje kolumn representerar en variabel.

Datatyp - Kategoriserar variabler baserat på deras egenskaper. De två huvudtyperna är kategoriska variabler och numeriska variabler. Kategoriska variabler innehåller kategorier, medan numeriska variabler består av siffror.

Distansmetrik - En formel som används för att beräkna distansen mellan två datapunkter, vilket kan ses som ett mått på deras likhet eller olikhet. Det finns olika distansmetriker och valet av metrik beror på användningsfall och egenskaperna hos variablerna.

Hyperparameter - Ett värde som behövs för att konfigurera en algoritm och som måste ställas in manuellt innan algoritmen påbörjas. Olika värden på hyperparametrar kan påverka prestanda och beteende hos algoritmen.

Klustra - Att gruppera datapunkter på ett sätt som placerar liknande datapunkter (enligt en specifik distansmetrik) i samma grupp eller kluster. Klustring är en vanlig teknik inom maskininlärning för att hitta naturliga grupper eller mönster i data.

Oövervakad maskininlärning - En typ av maskininlärning där algoritmen tränas på ostrukturerad data utan tillgång till något rätt svar. Målet är att upptäcka mönster eller strukturer i data. Klustring är ett exempel på en oövervakad maskininlärningsmetod.

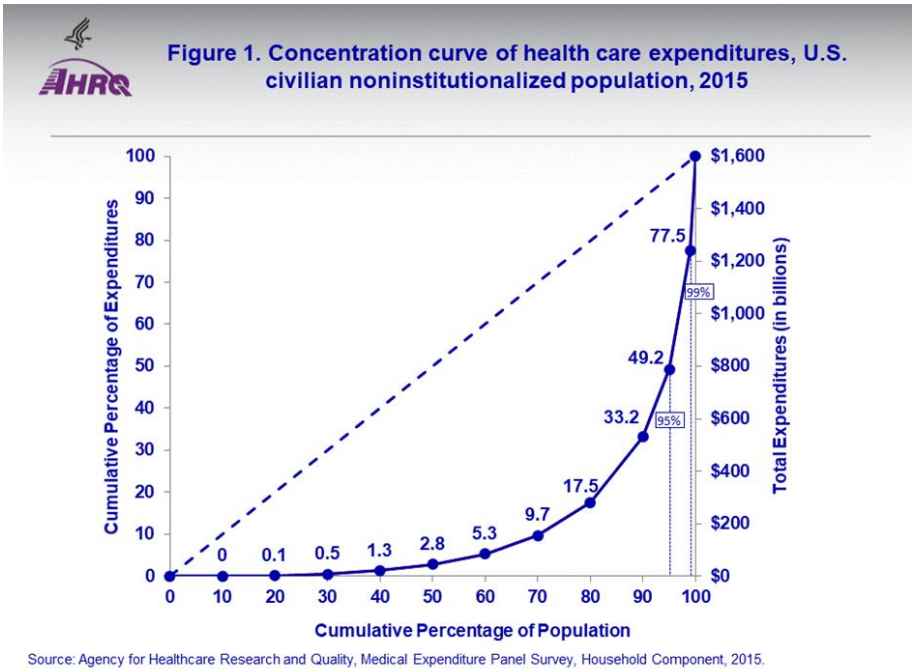
Parameter - En variabel eller inställning som används av en algoritm för att anpassa dess beteende eller prestanda under träning eller användning. Parametrar är vanligtvis inre variabler som justeras under träningsprocessen för att optimera modellens prestanda.

UMAP - Förkortning för Uniform Manifold Approximation and Projection, en datavetenskaplig teknik för visualisering av data med många variabler. UMAP är särskilt kraftfull då man vill bevara den inneboende strukturen mellan datapunkter men samtidigt representera datan i färre dimensioner (dvs. med färre variabler).

Variabel - En egenskap som beskriver olika aspekter hos datapunkter i ett dataset. Variabler kan vara antingen oberoende (prediktorer) eller beroende (responsvariabler) och representeras vanligtvis som kolumner i en datatabell. De kan vara av olika typer, inklusive numeriska, kategoriska eller ordinala, och används för att analysera och förstå data.

2 Introduktion

Det är väntat att kostnader för hälso- och sjukvård inte fördelas jämnt mellan olika patientgrupper. En överväldigande majoritet av de som söker vård gör det för besvär som kräver förhållandevis enkla åtgärder, vilka också medför låga kostnader. I stället används det mesta av budgeten för att finansiera behandlingar för allvarliga och/eller ovanliga tillstånd. Till exempel visar statistik från 2015 att 70 % av befolkningen stod för mindre än 10 % av USA:s totala sjukvårdskostnader under året, medan mer än hälften av samma kostnader var kopplade till de 5 % av patienterna som var mest kostsamma för sjukhusen (se figur 1). Flera OECD-länder observerar ett liknande mönster där de mest kostsamma patienterna svarar för mellan 40% och 60% av sjukhusens utgifter (källa: https://doi.org/10.1371/journal.pone.0217353). Även inom Sverige har tendensen påvisats, då specifikt för Region Västmanland som tagit fram siffrorna inom ramen för ett ännu pågående projekt.



Figur 1: Kostnadsfördelning med patienter indelade i centiler efter deras totala kostnader, rangordnade från lägst till högst kostnader.

Grundat i det uppstod en hypotes att en stor andel av hälso- och sjukvårdskonsumtionen i Västra Götalandsregionen (VGR) troligt fokuseras till en liten grupp patienter. Den gruppen benämns nedan ”kostnadsintensiva patienter”. Sammansättningen av gruppen kostnadsintensiva patienter är relativt outforskad såväl nationellt som internationellt, troligtvis på grund av gruppens inneboende medicinska komplexitet och det faktum att varje kostnadsintensiv patient ofta har flera diagnoser. Det är således svårt att studera kostnadsintensiva patienter med traditionella statistiska metoder som diagnosrelaterade grupper (DRG). Oövervakade maskininlärningsmetoder har emellertid potential att kunna hantera en högre komplexitet och därmed potential att kunna identifiera undergrupper bland kostnadsintensiva patienter. Sådana undergrupper kan bidra till förståelse kring vad som utmärker de kostnadsintensiva patienterna samt om det finns tidigare okända faktorer som driver deras kostnader.

Baserat på tidigare publicerade artiklar och genomförda samtal med kliniker och forskare antogs följande: 1) Det finns tydliga undergrupper inom gruppen av kostnadsintensiva patienter (till exempel patienter med sällsynta genetiska sjukdomar, patienter med olika typer av cancer, sköra äldre patienter) och 2) för vissa av dessa grupper förekommer det en ansenlig mängd så kallade "icke värdeskapande besök” (engelska: failure demand). Icke värdeskapande besök innebär att patienten erhåller vård eller undersökningar som egentligen inte motsvarar patientens behov (källa: doi:10.1108/IJHCQA-08-2017-0159). Att karaktärisera de kostnadsintensiva patienterna kan vara ett sätt att identifiera orsaker till icke värdeskapande besök eller undergrupper av patienter där icke värdeskapande besök är vanligt förekommande. Ökad förmåga att identifiera och avstyra dessa besök kan i sin tur

leda till både kostnadsbesparingar och förbättrade vårdprocesser för olika patientgrupper, vilket i längden kan ha positiv inverkan på de berörda patienternas livskvalitet.

Kompetenscentrum AI (Sahlgrenska Universitetssjukhuset) och enhet regional vårdanalys (Koncernstab Digitalisering) vid VGR initierade hösten 2023 en explorativ, data-driven studie för att 1) beskriva kostnadsfördelning mellan patienter i VGR vad gäller utförd specialistvård; 2) årsvis identifiera den kostnadsintensiva patientgruppen i VGR (patienter som tillhör de 5 % för sjukhusen mest kostsamma patienterna); samt 3) använda oövervakade maskininlärningsmetoder för att försöka skapa kluster och beskriva de kostnadsintensiva patienterna. Projektet finansierades delvis av Vinnova, genom deras paraplyprojekt ”informationsdriven vård”. Projektidén framkom efter inspiration från region Västmanland som genomfört en liknande analys, men då inte använt sig av maskininlärning.

2.2 Avgränsningar

Beslut togs om att i en första fas fokusera på patienter över 18 år i klusteranalysen, då mönster och vårdförlopp ofta skiljer sig mycket mellan barn och vuxna. Vidare har kostnader för primärvård helt exkluderats till följd av begränsad datatillgång, men det är samtidigt rimligt att anta att kostnader för primärvård är förhållandevis små i sammanhanget. Bristande datakvalitet orsakade av systembyten i regionen har också gjort det nödvändigt att utesluta data från år före 2015.

3 Metod

3.1 KPP-databasen

Kostnad Per Patient, KPP, är ett system för kostnadsfördelning inom sjukhusvård. Systemet täcker in kostnader för specialistvård vid regionens sjukhus eller där regionen köpt tjänster från andra vårdaktörer. VGRs KPP-data förvaltas av enhet regional vårdanalys (Koncernstab Digitalisering) och levereras sedan årligen till en nationell KPP-databas.

För att identifiera kostnadsintensiva patienter användes strukturerade databastabeller baserade på den regionala KPP-databasen. Regional vårdanalys ansvarade för att sammanställa pseudoanonymiserad KPP-data i ett passande format för projektet. Det huvudsakliga datasetet inkluderade information om alla vårdkontakter med sjukhus inom VGR från 2009 till 2022 (se Tabell 3 i Appendix för variabelbeskrivningar). Utöver det huvudsakliga datasetet erhölls separata tabeller för diagnoser och åtgärder, kopplade till det huvudsakliga datasetet med nyckeln *EpisodId*.

Under bearbetning av data upptäcktes att en påtagligt stor andel vårdkontakter hade ett saknat värde i variabeln huvuddiagnos. Enligt ansvariga för KPP i VGR kan det förklaras med att det inte är obligatoriskt att vid en vårdkontakt ange en diagnoskod om det inte rör sig om ett läkarbesök eller besök inom slutenvården. De saknade värdena verkar inte kunna härledas till någon speciell del av vårdsystemet, men påverkar förstås i viss mån projektet slutresultat.

3.2 Databearbetning

I ett inledande steg granskades det erhållna datasetet för att säkerställa att det matchade projektets inklusionskriterier och inte innehöll några uppenbara anomalier. Huvudsakliga åtgärder inkluderade följande:

- Då vårdkonsumtion och kliniska mönster inte är desamma för barn och vuxna exkluderades patienter under 18 år.

- Alla kostnader konverterades till 2021 års rådande penningvärde, för att möjliggöra jämförelser över tid. Utgångspunkten 2021 valdes baserat på datatillgång.
- För vissa år återfanns ett fåtal patienter med en negativ totalkostnad, mest troligt till följd av felregistreringar i databasen. Patienter med en negativ totalkostnad ett visst år exkluderades ur analysen det berörda året.
- I datasetet förekom den ogiltiga sjukhuskoden ”0”. Antalet rader med detta fel var väldigt få och exkluderades därför utan vidare undersökning.
- I de fall episodID inte var unikt för varje vårdkontakt, och därmed inte heller unikt för varje rad i datasetet, avlägsnades alla dubbletter till raden. Detta problem förekom i huvudsak bara för år tidigare än 2015, varför dessa år exkluderades helt från analysen.

Efter ovanstående bearbetning av datasetet delades patienterna in i grupper baserade på vårdkonsumtion. Indelningen gjordes genom att räkna ut den totala kostnaden för vårdkonsumtion per patient och år, varefter patienterna sorterades efter genererad kostnad så att varje års kostnadsintensiva patienter (de 5 % mest kostsamma patienterna) kunde identifieras.

3.3 Dataset för klustring

Klustringen utfördes årsvis och endast med data från den kostnadsintensiva patientgruppen. Ett dataset skapades som för varje år beskrev de kostnadsintensiva patienterna genom deskriptiva variabler som valts ut och bearbetats (se Tabell 1). En mer detaljerad beskrivning av alla variabeldefinitioner finns i filen ”README” i gitlab-kodbasen (https://git.vgregion.se/digital_foui/hnhc_patients)

Tabell 1: Beskrivning av de variabler som använts för klustring av den kostnadsintensiva patientgruppen.

Variabel	**Beskrivning**
age	Patientens ålder det aktuella året
sex	Patientens kön
was_hcp	Om patienten tillhörde den kostnadsintensiva gruppen föregående år
over_age_limit	Om patienten var över 65 år eller inte
total_diagnoses	Antal diagnoser satta det aktuella året
Aids	Om patienten diagnostiserats med aids eller HIV det aktuella eller föregående år
ami	Om patienten diagnostiserats med hjärtinfarkt det aktuella eller föregående år
canc	Om patienten diagnostiserats med cancer det aktuella eller föregående år
cevd	Om patienten diagnostiserats med cerebrovaskulär sjukdom (t ex stroke) det aktuella eller föregående år
chf	Om patienten diagnostiserats med hjärtsvikt det aktuella eller föregående år
copd	Om patienten diagnostiserats med KOL det aktuella eller föregående år
dementia	Om patienten diagnostiserats med demens det aktuella eller föregående år
diab	Om patienten diagnostiserats med diabetes (utan komplikationer) det aktuella eller föregående år

diabwc	Om patienten diagnostiserats med diabetes (med komplikationer) det aktuella eller föregående år
Hp	Om patienten diagnostiserats med hemiplegi eller paraplegi (förlamning) det aktuella eller föregående år
metacanc	Om patienten diagnostiserats med cancer med metastaser det aktuella eller föregående år
mld	Om patienten diagnostiserats med mild leversjukdom det aktuella eller föregående år
msld	Om patienten diagnostiserats med moderat eller allvarlig leversjukdom det aktuella eller föregående år
pud	Om patienten diagnostiserats med magsår det aktuella eller föregående år
pvd	Om patienten diagnostiserats med perifer kärlsjukdom det aktuella eller föregående år
rend	Om patienten diagnostiserats med njursjukdom det aktuella eller föregående år
rheumd	Om patienten diagnostiserats med reumatoid artrit det aktuella eller föregående år
mental_illness	Om patienten diagnostiserats med psykisk sjukdom det aktuella eller föregående år
mental_illness_substances	Om patienten diagnostiserats med psykisk sjukdom relaterad till drogmissbruk det aktuella eller föregående år
has_dialysis	Om patienten har fått dialys under det aktuella året
has_complications	Om patienten diagnostiserats med komplikationer till kirurgiska åtgärder och medicinsk vård under det aktuella året
has_injury	Om patienten diagnostiserats med skada eller förgiftning under det aktuella
has_covid	Om patienten diagnostiserats med covid-19 under det aktuella året
has_emergency	Om patienten besökt en akutmottagning under det aktuella året

3.4 Klustringsmetoder

Under projektets gång testades tre olika klustringsalgoritmer: KMeans, DBSCAN och HDBSCAN. Av de tre metoderna bedömdes att HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) hade bäst potential att spegla det givna datasetet. HDBSCAN är en utökning av den mer välkända klustringsmetoden DBSCAN och används liksom DBSCAN för att gruppera data baserat på densitet, det vill säga antalet datapunkter som ligger inom ett visst område. I jämförelse med DBSCAN sägs HDBSCAN dock vara bättre på att hantera datamängder där densiteten förväntas variera mellan kluster. HDBSCAN ska också vara mindre känslig för skiftningar i hyperparametrar (http://dx.doi.org/10.21105/joss.00205). HDBSCAN kan, förutom att klustra datapunkter också identifiera de datapunkter som inte följer datamängdens mönster, så kallat brus (engelska: noise). Som distansmetrik för klustringsalgoritmen användes Gower-distance. Med Gower-distansmetriken får man en samlad likhets- eller avståndsvärdering mellan varje par av datapunkter i en datamängd med variabler av olika datatyp, vilket väl passade de binära och kontinuerliga variabler som klustringen baserades på. Klustringen gjordes med hjälp av scikit-learns implementering av HDBSCAN.

3.5 Justering av hyperparametrar

För att optimera klustringen implementerades först en bayesiansk optimeringsmetod med hjälp av pythonpaketet scikit-optimize och funktionen *gp_minimize*. Funktionen applicerades på HDBSCANs hyperparametrar *min_cluster_size* och *min_samples*. *Min_cluster_size* anger det minsta antalet datapunkter som måste höra samman för att betraktas som ett kluster. *Min_samples* anger det minsta antalet datapunkter som ska finnas inom en viss distans ifrån en datapunkt för att datapunkten ska få betraktas som en så kallad kärnpunkt i HDBSCAN-algoritmen. Kärnpunkter kan ses som utgångspunkter för nya kluster. Hyperparametrarna optimerades initialt med avseende på valideringsmetriken silhouette score:

$$s = \frac{b - a}{\max(a, b)}$$

där a är den genomsnittliga distansen mellan en viss datapunkt i ett kluster och alla andra datapunkter I samma kluster, och b är den genomsnittliga distansen från en viss datapunkt i ett kluster till alla datapunkter i dess klusters närmaste kluster. Silhouette score kommer med andra ord premiera kluster där datapunkterna sitter tätt ihop och är väl avgränsade från angränsande kluster. Silhouette score tar dock inte hänsyn till hur många datapunkter som klassificeras som brus. Optimering av hyperparametrar baserat på denna valideringsmetrik resulterade därför i små och få kluster, och att en stor del av datapunkterna klassades som brus. Då syftet med klustringen var att få en överblick över vilka patientgrupper som finns inom den kostnadsintensiva gruppen var detta inte användbart. I stället anpassades hyperparametrarna manuellt för att få ett hanterbart antal kluster och begränsat med brus.

4 Resultat

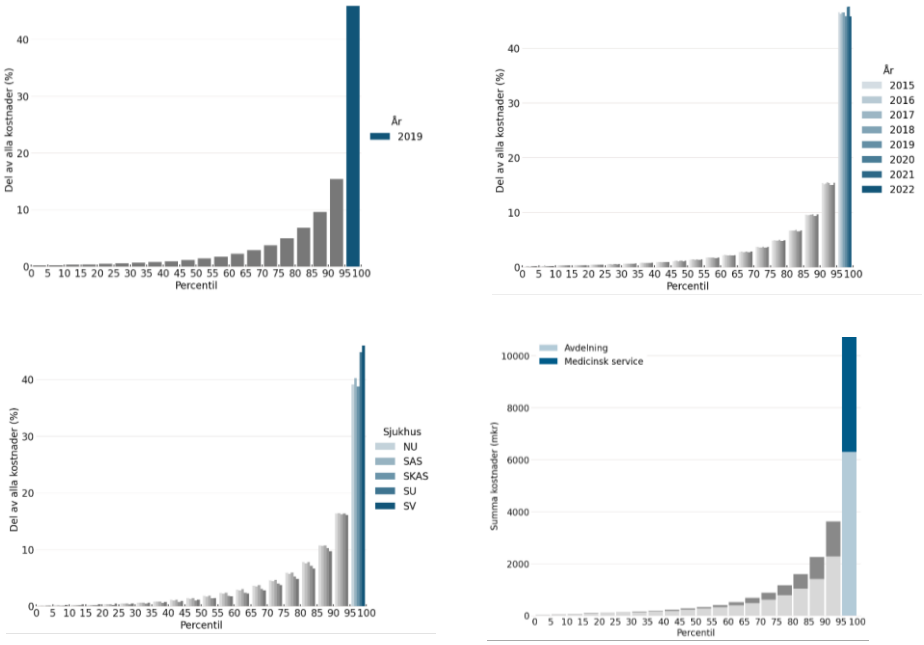
Initialt undersöktes de totala kostnaderna för sjukhusvård i VGR under året 2019. 727 934 patienter hade kontakt med VGRs sjukhus under 2019. Sammanlagt hade dessa patienter 3 436 737 vårdkontakter, varav 2 793 563 (81,3 %) vårdkontakter var planerade. Totalkostnaden för alla vårdkontakter var 27,3 miljarder kronor.

Då vi uteslutit patienter enligt 3.2 (till exempel de under 18 år) återstod 577 819 patienter och totalkostnaden 23,5 miljarder kronor.

Observera att alla kostnader i resultatet är justerade till penningvärdet för år 2021.

4.1 Identifiering av den kostnadsintensiva patientgruppen

Sortering av patienterna utefter kostnad bekräftade hypotesen att de 5% mest kostsamma patienterna står för ungefär 50% av de totala kostnaderna för sjukhusvård i VGR, se Figur 2a. Grafen beskriver samtliga patienter över 18 år som behövt sjukhusvård inom regionen år 2019. Patienterna rangordnas från lägst till högst kostnad i 20 lika stora grupper. År 2019 bestod gruppen kostnadsintensiva patienter över 18 år av 28 890 patienter, som tillsammans stod för 45,8% (10,8 miljarder kronor) av de totala kostnaderna för sjukhusvård. I Figur 2b visas samma indelning för åren 2015-2021, där det kan konstateras att mönstret är detsamma över tid. I Figur 2c syns istället de olika sjukhusens motsvarande indelningar. Även här ses mestadels samma mönster, men SU sticker ut med en något högre andel av budget som går till kostnadsintensiva patienter jämfört med övriga sjukhus. I Figur 2d visas kostnaderna uppdelat i avdelningskostnader och medicinsk service. För den kostnadsintensiva patientgruppen utgör kostnader för medicinsk service i snitt en större del av kostnaderna än för övriga patienter.

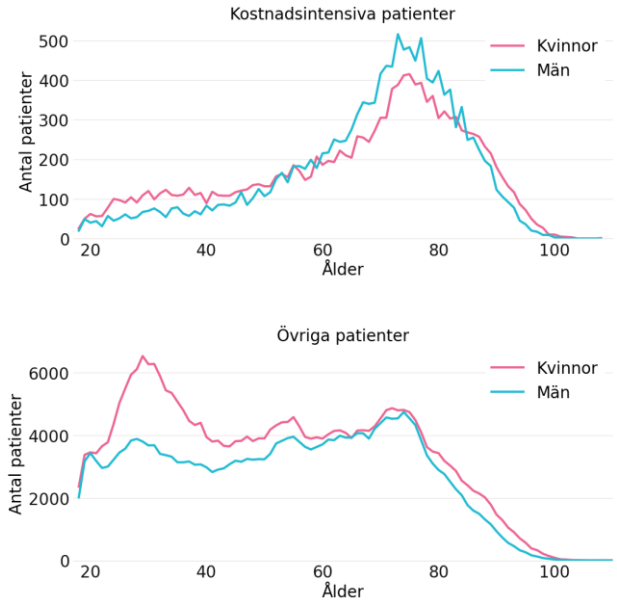


Figur 2: Fyra grafer som beskriver kostnadsfördelningen mellan olika percentiler av patienter, sorterade efter lägst till högst kostnad. I den övre vänstra grafen (a) syns kostnadsfördelningen för år 2019 för hela VGR. I den övre högra grafen (b) visas kostnaderna för 2015–2022. I den nedrevänstra grafen (c) visas igen kostnaderna för 2019, nu uppdelat per sjukhus. I den nedre högra grafen (d) visas kostnaderna för 2019 indelat i kostnadstyp.

4.2 Beskrivning av den kostnadsintensiva patientgruppen

Den kostnadsintensiva patientgruppen har en stor andel äldre patienter, som syns i Figur 3. Könsfördelningen bland de kostnadsintensiva patienterna verkar emellertid inte skilja sig nämnvärt från övriga patienter. I båda fallen syns till exempel att kvinnor i fertil ålder (20–40 år) är representerade i högre utsträckning än män i samma åldersspann. En skillnad syns i att den kostnadsintensiva gruppen har en ökad andel män från ca 50 års ålder och uppåt, medan det för övriga verkar vara en jämn fördelningen mellan män och kvinnor från ca 60 års ålder.

Vid kontroll av vanligast förekommande huvuddiagnoser upptäcktes skillnader mellan kostnadsintensiva patienter och övriga patienter. För kostnadsintensiva patienter var de vanligaste ICD-koderna Extrakorporeal dialys (Z491), Kemoterapeutisk behandling för tumör (Z511) och Opioidberoende (F112). För övriga patienter var motsvarande huvuddiagnos saknas, Opiodberoende (F112) och Psykisk störning ej specificerad på annat sätt (F999). Att huvuddiagnos saknas för en vårdkontakt kan förstås ha många orsaker, men kan innebära att vårdkontakten inte krävt att någon diagnos ställs (som vid kontakt med arbetsterapeut eller dietist). Tabeller över de 25 vanligaste huvuddiagnoserna för respektive grupp återfinns i appendix.

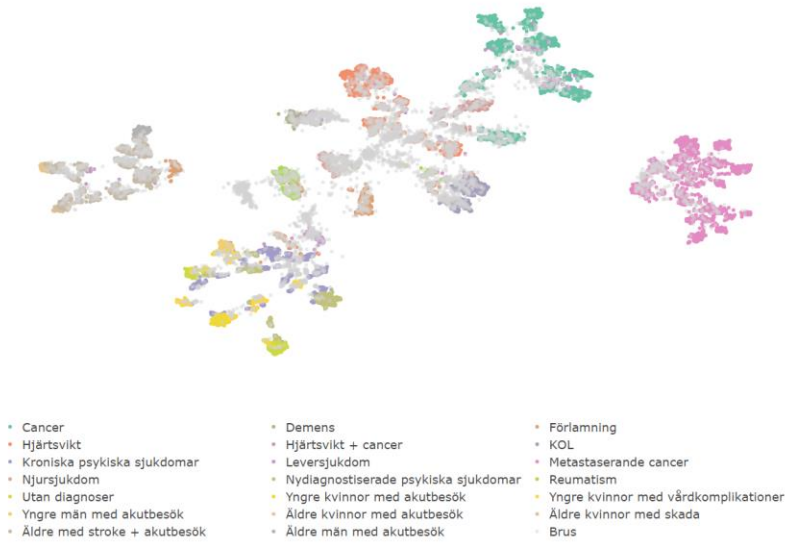


Figur 3: Åldersfördelning uppdelat på män och kvinnor. Den övre grafen visar de kostnadsintensiva patienterna, den nedre övriga.

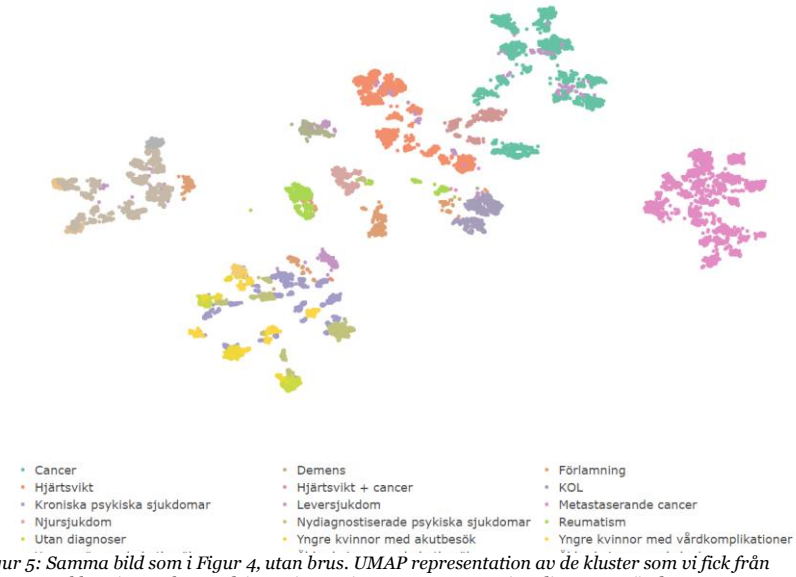
4.3 Klustrings-resultat

Efter HDBSCAN-klustring av den kostnadsintensiva patientgruppen för år 2019 erhölls cirka 20 kluster utöver de datapunkter som klassades som brus. Genom att granska information om patienterna som inkluderades i specifika kluster gjordes en kvalitativ utvärdering av klustren i syfte att identifiera deras kliniska profil. Motsvarande resultat för år 2021 finns i appendix.

För visualisering av klustringen användes den dimensionsreducerande algoritmen UMAP, på grund av dess förmåga att bevara globala strukturer i data. Det är viktigt att notera att UMAP har genomförts med distansmetriken Hamming, vilken alltså skiljer sig något från den Gower-distansmetrik som använts vid själva klustringen. Anledningen till att Hamming-distansen användes vid visualiseringen var främst att det gav implementeringsmässiga fördelar, och då visualiseringen inte ska ha någon påverkan på eventuella slutsatser som kan dras av resultaten bedömdes att det inte var av stor vikt att distansmetrikerna var identiska.



Figur 4: UMAP representation av de kluster som vi fick från HDBSCAN klustring av kostnadsintensiva patienter 2019. Hamming distans användes som distansmetrik för UMAP.



Figur 5: Samma bild som i Figur 4, utan brus. UMAP representation av de kluster som vi fick från HDBSCAN klustring av kostnadsintensiva patienter 2019. Hamming distans användes som distansmetrik för UMAP.

Tabell 2: Beskrivning av kluster bland de kostnadsintensiva patienterna år 2019. En illustration av klustren finns i ***Error! Reference source not found.****.*

Kluster	Antal patienter	Beskrivning
Brus	6101	En brus-kluster av patienter som algoritmen inte kunde placera i någon av de andra klustren. I genomsnitt är patienterna i denna grupp mer sjuka än resterande kluster.
Metastaserande cancer	4133	Ett stort kluster av patienter med metastatisk cancer. Klustret med högst komorbiditetsindex. Färre än hälften av dessa patienter har kontakt med sjukvården kommande år (2020), vilket kan indikera att en stor andel avlider innan dess.
Demens	378	Patienter i detta kluster är alla äldre med demenssjukdom. Nästan alla har besökt en akutmottagning. Färre än hälften av dessa patienter har kontakt med sjukvården kommande år (2020), vilket kan indikera att en stor andel avlider innan dess.
Förlamning	580	Majoriteten av patienterna har något typ av förlamning. Många har också en cerebrovaskulär sjukdom.
Leversjukdom	375	Patienter med leversjukdomar. Denna grupp har också en något högre grad av psykiska sjukdomar kopplade till psykoaktiva substanser.
Cancer	4041	Stort kluster av patienter med cancer, men inte metastaser.
Hjärtsvikt	2550	Äldre patienter med hjärtsvikt. Detta kluster har också högst andel hjärtinfarkter. Majoriteten har besökt en akutmottagning.
Hjärtsvikt + cancer	505	Äldre patienter som alla har cancer och hjärtsvikt. De har något högre komorbiditetsindex än andra kluster, och många olika diagnoser.
Njursjukdom	474	Patienter med njursjukdom och med många i behov av dialys.
KOL	1070	Patienter med KOL.
Reumatism	788	Patienter med reumatism, med en större andel kvinnor.
Äldre med stroke + akutbesök	2321	Äldre patienter med stroke, där majoriteten besökt akutmottagning.
Äldre män med akutbesök	345	Ett kluster med äldre män, som alla besökt akutmottagning.
Äldre kvinnor med skada	343	Ett kluster med äldre kvinnor, som alla besökt akutmottagning samt diagnostiserats med någon typ av skada.
Äldre kvinnor med akutbesök	350	Ett kluster med äldre kvinnor, som alla besökt akutmottagning.
Kroniska psykiska sjukdomar	1303	Yngre patienter med kroniska psykiska sjukdomar och flest antal diagnoser och besök per patient. Många av dessa var i den kostnadsintensiva gruppen även föregående år.
Yngre kvinnor med akutbesök	596	Yngre kvinnor med akutbesök.
Yngre kvinnor med vårdkomplikationer	642	Yngre kvinnor, där majoriteten besökt akutmottagning. Det är också klustret med högst andel vårdkomplikationer.

Nydiagnostiserade psykiska sjukdomar	875	Yngre patienter med nydiagnostiserade psykiska sjukdomar.
Yngre män med akutbesök	308	Ett kluster bestående av yngre män, som alla besökt akutmottagning.
Utan diagnoser	812	En patientgrupp utan några av de diagnoser vi inkluderade i klustringen. Kan var en grupp med patienter med funktionella sjukdomar och därmed hittas inga matchande diagnoser.

5 Diskussion

5.1 Tolkning av resultat

Oövervakad maskininlärning är alltid till viss del subjektiv och det är svårt att bedöma resultatens relevans i och med att ett facit saknas. I det här fallet har variabler selekterats och tolkningar gjorts utifrån projektdeltagarnas egen förförståelse, samt vissa medicinska inspel från läkare som inriktat sig på sällsynta diagnoser.

Vidare kräver HDBSCAN-algoritmen inställningar av hyperparametrar såsom *min_cluster_size* och *min_samples*. Valet av dessa hyperparametrar påverkar i stor utsträckning de slutliga klustren och togs här fram genom manuella justeringar för att uppnå ett hanterbart antal kluster.

Det kan trots detta fastslås att det är möjligt att använda HDBSCAN för att skapa kluster av patientdata och identifiera meningsfulla undergrupper. Undergrupperna tycks spegla olika patientprofiler som kan vara relevanta ur ett kliniskt- eller forskningsperspektiv. För en grundligare och mer objektiv analys skulle ett specifikt användningsfall kunna hjälpa till att vägleda i hur metoden kan utvecklas. Ett specifikt användningsfall skulle troligtvis möjliggöra mer precis tolkning av de undergrupper som framträder och kunna tjäna som grund för ytterligare forskning eller klinisk tillämpning.

5.2 Framtida arbete

Vi ser att det finns ett flertal sätt på vilka man kan arbeta vidare med den kostnadsintensiva patientgruppen, vilket i framtiden kan ge den här analysen klinisk betydelse. Då de kostnadsintensiva patienterna tar stora resurser i anspråk (ekonomiska, personella och platsrelaterade) kan man tänka sig att riktade insatser mot denna patientgrupp skulle göra stor skillnad för förbättring och effektivisering av vården. Förhoppningar finns om att en vidareutveckling av analysen som här påbörjats skulle kunna falla ut i förebyggande åtgärder och optimeringar av vårdprocesser som gynnar både såväl enskilda patienter som regionen i sin helhet.

Ett sätt att få reda på mer om den kostnadsintensiva gruppen vore att låta fler datakällor ligga till grund för analysen. De variabler som använts här utgör bara en väldigt liten del av den information som skulle kunna samlas om patienterna. För kunskap om en patients hela sjukvårdsinteraktion hade till exempel också krävts data från primärvården.

En annan möjlig väg framåt vore att separat skapa kluster även för den grupp patienter som inte är kostnadsintensiva för att sedan se om det finns kluster i de två patientgrupperna där patienterna, trots kostnadsskillnaderna, delar medicinska egenskaper. Det skulle kunna göra det möjligt att identifiera faktorer som avgör huruvida en patient med en viss klinisk profil blir kostnadsintensiv. Kunskap om dessa faktorer skulle innebära, i den mån de är påverkbara, möjlighet att minimera antalet patienter som blir kostnadsintensiva.

Slutligen är det även av intresse att undersöka närmre vilka patienter som återkommer i den kostnadsintensiva gruppen år efter år, för att se om vården för denna specifika undergrupp kan förbättras.

5.3 Slutsats

Analysen har påvisat och gett en initial beskrivning av en kostnadsintensiv patientgrupp i VGR, definierad som 5 % av det totala antalet patienter som ensamma medför cirka 50 % av sjukhusens kostnader. Deskriptiva analyser visar på egenskaper som särskiljer patienter i den kostnadsintensiva gruppen. Vidare har vissa mer komplexa medicinska profiler hos kostnadsintensiva patienter kunnat identifieras genom klusteranalys. Emellertid kvarstår arbete innan resultaten kan få klinisk relevans, ett arbete som skulle underlättas av att vidare analyser grundas i en specifik och välavgränsad frågeställning.

6 Bra att veta

6.1 Kontakt

Arbetet har utförts som ett samarbete mellan Kompetenscentrum AI på Sahlgrenska Universitetssjukhuset (SU) och enhet Regional Vårdanalys på Koncernstab Digitalisering (KsD) inom Västra Götalands Regionen.

Viktoria Karlsson (KsD), data scientist - viktoria.c.karlsson@vgregion.se

Tove Mathiasson (KsD), strateg - tove.mathiasson@vgregion.se

Anna Rosén (SU), data scientist -

Lisa Sjöblom (SU), data scientist - lisa.l.sjoblom@vgregion.se

Juulia Suvilehto (SU), överdatascientist – Juulia.suvilehto@vgregion.se

6.2 Kod

Koden som skapats för projektet finns på VGRs GitLab och kan klonas genom:

```
Git clone https://git.vgregion.se/digital_foui/hnhc_patients.git
```

7 Appendix

7.1 Beskrivning av databastabell

Tabell 3: Beskrivning av de variabler ur KPP-databasen som använts.

Kolumn	Möjliga värden	Beskrivning
År		År
Vårdtyp	OV/SV	Öppenvård/slutenvård
PersonKod		Personnummer, 12 siffror
IndividID		Anonymt ID. Kunde ej användas då det förekom felaktigheter.
EpisodId		ID för raden, alltså för vårdkontakten.
Kon	NULL/0/1/2	Kön, där 2 betyder kvinna
Alder		Ålder vid vårdkontakten

Hemort		Kod för hemorten, fyra siffror
LTGrupp	VGR/Halland/Övr	Region som patienten bor i
Sjukhus		Kod för vilket sjukhus
Sjukhuskort		Kod eller förkortning för sjukhus
MVOKod		Kod för medicinskt verksamhetsområde
InDatum		Datum för inskrivning
UtDatum		Datum för utskrivning
V_tid		Antal dagars vårdtid
PlaneradKod	J/N	Om vården var planerad eller akut
InSätt		Kod för hur patienten inkom till sjukhuset
UtSätt		Kod för hur patienten lämnade sjukhuset
Vårdnivå	L/N/R/P/Q/0	Länssjukvård, Regional Högspecialiserad vård, nationell högspecialiserad vård, etc
EkonomisktOmradeKod	0/Psykiatri/Somatik	Område där kontakten skedde
DRG		Kod för DRG som kontakten tillhör
YO1		Yttre orsak som skrivs i kombination med ICD-10 kod. Skrivs i kombination med skador och komplikationer
YO2		Se ovan
YO3		Se ovan
TotKost		Total kostnad för kontakten
TotKost2023		Total kostnad justerad till penningvärdet 2023
TotKost2022		Total kostnad justerad till penningvärdet 2022
TotKost2021		Total kostnad justerad till penningvärdet 2021
SumAvd		Kostnad för avdelning
SumMs		Kostnad för medicinsk service
Rtg		Kostnad för röntgen
OpKost		Kostnad för operation
IVA		Kostnad för IVA
Mtrl		Kostnad för material
Lakemedel		Kostnad för läkemedel
KemLab		Kostnad för kemlabb
KlinFys		Kostnad för klinisk fysiologi
Patologi		Kostnad för patologi
BaktLab		Kostnad för BaktLab
Blod		Kostnad för blod
ÖvrigtLab		Kostnad för övriga labb
Sjukgymnast		Kostnad för sjukgymnastik
Arbetsterapeut		Kostnad för arbetsterapeut
ÖvrigaTjänster		Kostnad för övriga tjänster
Akut		Kostnad för akutmottagning
KVA		KVÅ, klassifikation av vårdåtgärder
DRGKortNamn2022		Kort textbeskrivning av DGR-koder
Vikt2022		Vikt för kostnadsomvandlingen för 2022
Yfgräns2022		Ytterfallsgräns år 2022
Yfall	I/Y	Ytterfall är vårdkontakter som kostar betydligt mer än genomsnittet, baserat på DRG

7.2 Vanligaste diagnoskoderna 2019

Tabell 4: De vanligaste huvuddiagnoserna för den kostnadsintensiva patientgruppen år 2019. Kolumnen andel indikerar andelen av alla vårdkontakter inom denna grupp som hade denna huvuddiagnos.

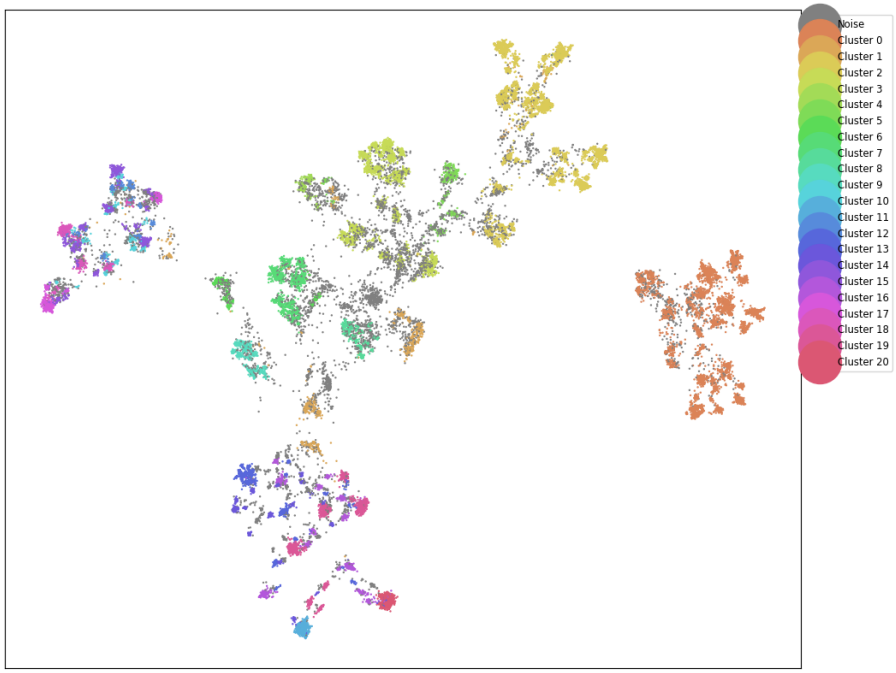
	ICD-kod	Antal	Andel
1	Z491	54048	8,49 %
2	Z511	36524	5,74 %
3	F112	32329	5,08 %
4	Z510	27509	4,32 %
5	Okänd	16439	2,58 %
6	F603	10563	1,66 %
7	I509	8433	1,33 %
8	C509	7695	1,21 %
9	Z090	6033	0,95 %
10	F431	4838	0,76 %
11	C900	4788	0,75 %
12	F192	4747	0,75 %
13	F999	4485	0,70 %
14	F200	4416	0,69 %
15	F299	4214	0,66 %
16	C619	3822	0,60 %
17	F259	3694	0,58 %
18	F209	3690	0,58 %
19	F900B	3624	0,57 %
20	C209	3549	0,56 %
21	Z080	3490	0,55 %
22	F500	3362	0,53 %
23	N185	3245	0,51 %
24	C349C	3220	0,51 %
25	F412	3174	0,50 %

Tabell 5: De vanligaste huvuddiagnoserna för den patienter utanför den kostnadsintensiva gruppen år 2019. Kolumnen andel indikerar andelen av alla vårdkontakter inom denna grupp som hade denna huvuddiagnos.

	ICD-kod	**Antal**	**Andel**
1	Okänd	83138	3,73 %
2	F112	37155	1,67 %
3	F999	31230	1,40 %
4	F900B	27927	1,25 %
5	Z090	27046	1,21 %
6	Z510	24416	1,09 %
7	F102	21920	0,98 %
8	F412	20113	0,90 %
9	Z392	19425	0,87 %
10	R104X	19143	0,86 %

11	Z363	16888	0,76 %
12	R074	16842	0,76 %
13	F431	16775	0,75 %
14	C619	16179	0,73 %
15	I509	15515	0,70 %
16	F603	15416	0,69 %
17	H353B	14632	0,66 %
18	Z032	14612	0,66 %
19	E109	14580	0,65 %
20	G473	14469	0,65 %
21	O800A	14006	0,63 %
22	Z768	12991	0,58 %
23	H259	12982	0,58 %
24	Z349	12870	0,58 %
25	Z369	12436	0,56 %

7.3 Klustringsresultat för år 2021



Figur 6: UMAP representation av de kluster som vi fick från HDBSCAN klustring. Hamming distans användes som distansmetrik för UMAP.

Tabell 6: Beskrivning av kluster bland de kostnadsintensiva patienterna år 2019. En illustration av klustren finns i Figur 6.

Kluster	Antal patienter	Beskrivning
-1	6354	Brus
0	4280	Högst antal komorbiditeter, i genomsnitt över 50 diagnoser inklusive metastaserande cancer där 1/3 varit hcp föregående år, denna grupp har den näst högsta patologikostnaden.
1	865	En grupp som innehåller flest hcp med leversjukdomar, aids, cerebrovaskulära sjukdomar och nästan hälften har hemiplegi eller paraplegi.
2	4054	Detta är ren stor grupp som innehåller patienter med cancer, de har den högsta associerade patologikostnaden.
3	2348	Denna grupp har färre totala besök jämfört med de andra grupperna, medianåldern är över 80 och alla har hjärtsvikt med 25% akut hjärtinfarkt, nästan 90 % av patienterna hade akutrelaterade besök. Har den högsta mediankostnaden för klinfys-relaterade utgifter.
4	314	Denna grupp har färre än 5 besök i genomsnitt per person, gruppen är äldst och alla har demenssjukdom. Nästan alla har haft akuta besök och psykisk sjukdom. Only half of these patients have costs associated with the hospital year 2022.
5	510	Denna grupp har en hög samtidig sjuklighet, är äldre och har hjärtsvikt med 20 % akut hjärtinfarkt, de har även cancer, och de flesta har varit på sjukhuset med ett akutbesök.
6	313	Patienter med diabetes inklusive komplikationer där många har haft akuta besök. De har bland de högsta kostnaderna för bakteriologilabb
7	1174	En grupp där patienterna har kol, de har färre besök i genomsnitt men en hög andel akuta besök
8	437	Denna grupp har njursjukdomar och en tredjedel av den här patientgruppen har fått dialys. De har också den högsta labb-associerade kostnaderna.
9	547	Alla patienter i denna grupp har reumatoid artrit.
10	343	Patienter i denna grupp är äldre där många har perifera kärlsjukdomar och komplikationer.
11	397	En ung patientgrupp som enbart består av kvinnor med akuta besök.
12	344	En grupp med enbart äldre patienter. En tredjedel har diagnostiserats med hjärtinfarkt och många i gruppen har haft komplikationer.
13	678	Den här gruppen innehåller patienter som har många diagnoser och många besök, de är unga och har någon form av psykiska besvär.
14	584	Denna grupp liknar kluster 16 men har ännu fler diagnoser och besök i genomsnitt. Av dessa var hälften också kostnadsintensiva patienter år 2020. En del av denna grupp har också diagnostiserats med psykisk sjukdom relaterad till drogmissbruk
15	1054	Denna grupp är äldre med få besök men en hög andel akuta samt 20% haft covid.
16	918	En medelåldersgrupp där det finns en hög andel med komplikationer

17	599	En äldre grupp som har skador som har lett till att de kommit in till akuten. De har annars få besök.
18	693	Äldre kvinnor som har sökt akutvård
19	1164	Medelålders män där 20% har haft covid
20	490	Yngre kvinnor utan akuta besvär