Date: 14th June 2017

GP Data Implementation Project

Sizing Extracts by complexity and capacity

**Document Management**

**Revision History**

|  |  |  |
| --- | --- | --- |
| Version | Date | Summary of Changes |
| 0.4 | 20/04/2017 | Amended to clarify departed patient numbers |

**Contents**

Contents

[1 Introduction 5](#_Toc479860288)

[2 Overview 5](#_Toc479860289)

[3 Complexity 6](#_Toc479860290)

[3.1 BR Complexity model 6](#_Toc479860291)

[3.2 SQL Complexity model 6](#_Toc479860292)

[3.3 Complexity banding 6](#_Toc479860293)

[4 Volume 8](#_Toc479860294)

[4.1 Volume model 8](#_Toc479860295)

[4.2 Volume banding 9](#_Toc479860296)

[5 Capacity 9](#_Toc479860297)

[5.1 Capacity units 9](#_Toc479860298)

[5.2 Capacity limits 9](#_Toc479860299)

[6 Appendix A – Existing extract Complexity estimates 12](#_Toc479860300)

[7 Appendix B – Existing extract Volume estimates 13](#_Toc479860301)

[8 Appendix C – Existing extract Capacity Units estimates 14](#_Toc479860302)

[9 Appendix D – Estimate of Capacity usage 15](#_Toc479860303)

[9.1 Existing extracts 15](#_Toc479860304)

[9.2 Three year forward view 16](#_Toc479860305)

# Introduction

Data extracts under GPES were sized according to a complexity model referred to as “Dimensions” and “Extract Capacity Units” (ECU’s).

Dimensions are largely based on a count of the Read Codes used in the Extract Specification, and as such are not considered a good reflection of the development effort needed to build the extract.

ECU’s are based on the calculated complexity of the Quality Outcomes Framework (QOF) extract, and are not considered a good reflection of the capacity usage of running the extract on the GPSS infrastructure.

Additionally, Dimensions are time-consuming to produce and rely on a completed ER Pack for their foundation.

For “GPES Uplift”, we are formulating new calculations for complexity and capacity that better represent the effort to build and system usage, respectively.

There is a risk that GP System Suppliers (GPSS) do not consider these as an ideal representation owing to differences in how the various suppliers have designed their extract query solutions. However, it has been designed based on feedback from the GPSS, and furthermore understanding what the complexity and capacity scores are for all current GPES extracts means we should be able to work with the GPSS to agree a baseline for what small/ medium/ large extracts look like at the point of their elaboration.

# Overview

We need to size extracts in two separate but interconnected ways.

1. Complexity
* Based on the content of the business rules, an estimate of how much work will be required by the GPSS to complete the build.
1. Capacity
* Capacity will be based on a combination of the Complexity calculation above and the anticipated relative data volumes that will be returned as a result of running the query.
* The rationale is that an extract that is complex to build is likely to exert a greater CPU load on the GPSS infrastructure, while a simple extract that is simple to run but returns large volumes of data is likely to exert a greater bandwidth load on the GPSS infrastructure.

The overall calculations need to reflect that there are essentially two kinds of data extracts that are requested through GPES, and that the complexity and capacity need to be broadly understood during the elaboration to determine timescales for their development and impact on available extract capacity on the GPSS infrastructure:

1. Aggregate
* Aggregate extracts return small volumes of data, as they consist of counts summarised to a practice level. However, they can involve a large amount of calculation (i.e. complexity) to run.
* QOF is a good example of a complex aggregate extract.
1. Patient level (sometimes referred to as PID-level)
* Patient level extracts can have complex rules, as per aggregate extracts. However, they can also result in the need to return high volumes of data.
* Overall capacity calculation should therefore consider both the complexity and expected data volumes.
* Diabetic Retinopathy (GP2 DRS) is a good example of a PID level extract.

# Complexity

There are two models for Complexity calculation, the main one based on the Business Rules (BR). In those rare cases where BR’s aren’t produced for an extract there is a second model based on Structured Query Language (SQL) specifications.

## BR Complexity model

The following algorithm is used to determine the complexity from the BR’s:

**Complexity Points** = (Number of “AND” + Number of “OR” + Number of Rules)

**Complexity Score** = ROUND(SQRT(Complexity Points))

## SQL Complexity model

The following algorithm is used to determine the complexity from the SQL:

**Complexity Points** = (Number of “ON” + Number of “WHERE” + Number of “AND” + Number of “OR”)

**Complexity Score** = ROUND(SQRT(Complexity Score))

Note that this method is influenced by the efficiency of the SQL that has been produced to define the extract logic. It can mean that more complex extracts appear simpler if the SQL is written more efficiently, and vice versa.

## Complexity banding

Complexity is then banded into “t-shirt” sizes, as per Table 3.3.1:

|  |  |
| --- | --- |
| **Complexity Score** | **Complexity** |
| 0-10 | Low |
| 11-20 | Medium |
| 21-40 | High |
| 41+ | Very high |

Table 3.3.1 – Complexity banding

Appendix A contains the complexity calculations for all extracts currently defined through GPES.

# Volume

A model is required to enable quantification of the average data volumes expected to be returned in PID-level extracts.

## Volume model

The algorithm in Table 4.1.1 shows the criteria we’re proposing to use to calculate the size of an average data file:

|  |  |  |
| --- | --- | --- |
| **Ref.** | **Clinical Codes** | **Notes** |
| CC1 | Number of codes possible to be returned in extract | This expanded cluster list count would come from Primary Care Domain (PCD)= CC2 + CC3 + CC4 |
| CC2 | Number of codes which return 'all' | Example – return all dementia records within specified code range |
| CC3 | Number of codes which return only once | Example – return the latest dementia record within specified code range |
| CC4 | Number of codes returned 'multiple' times but less than 'all' | Example – return the last 3 dementia records within specified code range |
|  |  |  |
|  | **Size Parameters**  |  |
| SP1 | Registered Practice list Size | Value fixed at average practice size of 6000 registered patients |
| SP2 | Percentage of Registered Practice List Size selected | Assumption based on knowledge of the dataset to be returned. Example: Diabetic patients are approximately 6% of the practice list |
| SP3 | Number of Non-Registered (e.g. Deceased/Left) Patients to be included | Some extracts will return records for departed patients. This is the *total number* of departed patients expected per extract. |
| SP4 | Assumed Mean Number of Codes/ Referrals/ Encounters/ Demographics per Patient | This will be an assumption based on understanding the implications of CC2, CC3 and CC4, and could be tested with the GPSS who may have better data to inform this parameter. |
| SP5 | File size for a single patient in KB | Sample XML file is manually created by NHS Digital using parameters SP4 and any other attributes from the four table which are to be included in the extract. This value is the file size of the sample file. |
|  |  |  |
|  | **Calculated Size** |  |
| CS1 | Number of Registered Patients in extract | = (SP1 / 100) \* SP2 |
| CS2 | Total Number of Patients | = CS1 + SP3 |
| CS3 | File size for the full cohort in KB | = SP5 \* CS2 |
|  |   |   |
| Volume in MB | File size for Practice in MB | = CS3 / 1024 |

Table 4.1.1 – Volume algorithm

See Appendix B for example calculations of Volume for existing PID-level extracts.

## Volume banding

Based on the Volume in MB, Capacity is then banded as per Table 4.2.1:

|  |  |
| --- | --- |
| **Volume in MB** | **Volume** |
| 0-20 | Low |
| 21-50 | Medium |
| 51-300 | High |
| 301+ | Very high |

Table 4.2.1 – Volume banding

Note that aggregate extracts can be safely assumed as <20MB, and so will always be included in the Low banding.

# Capacity

A model is required to combine the Complexity and Volume bands to arrive at a Capacity unit.

## Capacity units

Capacity usage estimation will be a matrix of complexity and volume bandings to arrive at a final Capacity score. Table 5.1.1 shows the points value within the Complexity/ Volume matrix.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | ComplexityVolume | Low | Medium | High | Very High |
| PID-level | Aggregate | Low | 1 | 2 | 4 | 6 |
|  | Medium | 2 | 4 | 6 | 12 |
| High | 4 | 6 | 12 | Negotiable |
| Very High | 8 | 12 | Negotiable | Negotiable |

Table 5.1.1 – Capacity units matrix

## Capacity limits

GPES Uplift provides a new scheduling model, which can be seen in Diagram 5.2.1. We must agree maximum Capacity units per day to enable the GPSS to cost their solution.

The schedule is based on the currently known extract for FY 17/18, with each DPW “Bucket” occupying set dates in each month.



Diagram 5.2.1 – GPES Uplift schedule

Assumption is for GPSS to provide a Maximum Capacity of 4 units for each day within each Data Processing Window (DPW), resulting in total units per bucket as per Table 5.2.2.

The capacity available to run extracts will be determined by the combination of units per day and the total number of days in the bucket – i.e. available capacity can be spread across the days within an individual bucket, but cannot span several buckets.

|  |  |  |
| --- | --- | --- |
| **Bucket** | **Number of DPW days** | **Total number of available units** |
| 1 | 2 | 8 |
| 2 | 3 | 12 |
| 3 | 5 | 20 |
| 4 | 3 | 12 |
| 5 | 3 | 12 |

Table 5.2.2 – Available Capacity Units per Bucket

Note that these parameters limit the Volume and Complexity that we can ask the GPSS to provide via the “Aggregate” and “Bespoke” feeds, as highlighted in red in Table 5.2.3, as there are no buckets containing more than 20 units. There are currently no extracts containing this level of Complexity or Volume.

Also, once current extracts are taken into consideration (see Appendix D), there will not be available headroom to run new extracts of the Complexity/ Volumes highlighted in amber in table 5.2.3.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | ComplexityVolume | Low | Medium | High | Very High |
| PID-level | Aggregate | Low | 1 | 2 | 4 | 6 |
|  | Medium | 2 | 4 | 6 | 12 |
| High | 4 | 6 | 12 | **Negotiable** |
| Very High | 8 | 12 | **Negotiable** | **Negotiable** |

Table 5.2.3 – Capacity Limitations on overall Complexity/ Volumes that GPES Uplift can accommodate

# Appendix A – Existing extract Complexity estimates

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Extract** | **PID?** | **Number of Rules** | **AND** | **OR** | **ON** | **WHERE** | **Complexity points** | **Complexity Score** | **Complexity** |
| **Alcohol** | N | 59 | 53 | 10 | 0 | 2 | 124 | 11 | Medium |
| **Childhood seasonal Flu** | N | 43 | 63 | 2 | 0 | 0 | 108 | 10 | Low |
| **Dementia** | N | 83 | 18 | 12 | 0 | 9 | 122 | 11 | Medium |
| **GP2DRS** | Y | 0 | 27 | 0 | 16 | 41 | 84 | 9 | Low \* |
| **GP2DRS MP** | Y | 0 | 6 | 4 | 15 | 18 | 43 | 7 | Low \* |
| **GP Insight Data** | Y | 146 | 94 | 35 | 0 | 0 | 275 | 17 | Medium |
| **INLIQ** | N | 418 | 124 | 55 | 0 | 2 | 599 | 24 | High |
| **LD** | N | 35 | 6 | 3 | 0 | 0 | 44 | 7 | Low |
| **LDO** | N | 310 | 122 | 81 | 0 | 0 | 513 | 23 | High |
| **Men ACWY** | N | 26 | 20 | 2 | 0 | 0 | 48 | 7 | Low |
| **Men B** | N | 69 | 65 | 22 | 0 | 2 | 158 | 13 | Medium |
| **Named accountable GP** | N | 24 | 10 | 2 | 0 | 0 | 36 | 6 | Low |
| **NHS Health Checks** | Y | 201 | 113 | 2 | 0 | 4 | 320 | 18 | Medium |
| **Pertussis** | N | 19 | 12 | 3 | 0 | 0 | 34 | 6 | Low |
| **Pneumoccocal** | N | 60 | 56 | 77 | 0 | 0 | 193 | 14 | Medium |
| **POM** | Y/N | 0 | 27 | 0 | 16 | 41 | 84 | 9 | Low \* |
| **QOF** | N | 868 | 331 | 104 | 0 | 13 | 1316 | 36 | High |
| **Rotavirus** | N | 24 | 12 | 0 | 0 | 0 | 36 | 6 | Low |
| **Seasonal Flu** | N | 142 | 155 | 101 | 0 | 1 | 399 | 20 | Medium |
| **Shingles catchup** | N | 29 | 22 | 4 | 0 | 0 | 55 | 7 | Low |
| **Shingles Routine** | N | 29 | 22 | 4 | 0 | 0 | 55 | 7 | Low |

Table 6.1 – Existing extract Complexity estimates

\* Note that POM, GP2 DRS and GP2 DRS MP have been calculated using the SQL Model, which is not so reflective of actual complexity.

# Appendix B – Existing extract Volume estimates

Refer to section 4.1 for underlying descriptions and notes regarding each category.



Table 7.1 – Existing extract Volume estimates

# Appendix C – Existing extract Capacity Units estimates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Extract** | **PID?** | **Complexity** | **Volume** | **Capacity** |
| **Alcohol** | N | Medium | Low | 2 |
| **Childhood seasonal Flu** | N | Low | Low | 1 |
| **Dementia** | N | Medium | Low | 2 |
| **GP2DRS** | Y | Low | Low | 1 |
| **GP2DRS MP** | Y | Low | Low | 1 |
| **GP Insight Data** | Y | Medium | Low | 2 |
| **INLIQ** | N | High | Low | 4 |
| **LD** | N | Low | Low | 1 |
| **LDO** | N | High | Low | 4 |
| **Men ACWY** | N | Low | Low | 1 |
| **Men B** | N | Medium | Low | 2 |
| **Named accountable GP** | N | Low | Low | 1 |
| **NHS Health Checks** | Y | Medium | High | 6 |
| **Pertussis** | N | Low | Low | 1 |
| **Pneumoccocal** | N | Medium | Low | 2 |
| **POM** | Y/N | Low | Low | 1 |
| **QOF** | N | High | Low | 4 |
| **Rotavirus** | N | Low | Low | 1 |
| **Seasonal Flu** | N | Medium | Low | 2 |
| **Shingles catchup** | N | Low | Low | 1 |
| **Shingles Routine** | N | Low | Low | 1 |

Table 8.1 – Existing extract Capacity Units estimates

# Appendix D – Estimate of Capacity usage

## Existing extracts



Table 9.1.1 – Estimate of Capacity Usage for existing extracts

Notes:

* This leaves capacity for 4-7 additional payment extracts in “Bucket 3” – the bucket for payment extracts.
* Those extracts which run quarterly/ off-set quarterly, only impact on the capacity during the months in which they are run

## Three year forward view

The view of upcoming extracts up to and including FY20/21:



Table 9.2.1 – Estimate of Capacity Usage over next three Financial Years