GIST NICE submission (review TA196, ID696)
Manufacturer Clarification Response; 13th March 2014

Section A: Clarification on clinical data

A1. Priority question: Page 62 refers to a HR (0.43 [95% CI 0.30 to 0.62]) for the high risk group. Please specify what the time period is; is it four years? (p63 top)

This HR represents the overall HR at the 5 year follow up. Please see page 207 of 'SSG study Heintz et al full CSR_sti571bfi03--legacy-clinical-study-report.pdf'.

A2. Priority question: Please clarify which population the overall survival data on page 70 if for (median follow up of 60.2 months for the ACOSOG Z9001 trial). Is it the ITT population or Miettinen high risk population?

This relates to the ITT population. Please see page 73 of 'ACOSOG 5 yrs sti571bus89--legacy-clinical-study-report 5-yr update.pdf'.

Please also clarify "overall population" in the same paragraph; is it the primary analysis population?

Yes, it is the primary analysis population. The term "overall population" in this paragraph should be changed so that the sentence states: "As for the primary analysis, these data...."

A3. Figure B-2 shows there were 2 rounds of title/abstract screening. Please clarify why there were 2 rounds.

Electronic screening was performed in two stages with the first round focussing on ensuring that all references meeting the inclusion criteria were included while the second round focussed on excluding references meeting the exclusion criteria. The latter were refined in the light of the findings of the first round and included addition of exclusion based on the number of patients and exclusion of GIST subgroups.

A4. Please provide a reference list for the 6 full papers that were excluded (Figure B-2)

The following are the 6 references that were excluded. Please note that two were excluded because they were duplicates (rather than for study design). Below we provide an updated version of the PRISMA which clarifies that two references were excluded because they were duplicates rather than for design.

Excluded for study design:
Ozaka, M., M. Ogura, et al. (2011). "The optimum time for surgery after receiving imatinib: A retrospective analysis of GIST patients achieving a partial response (Japan)." Japanese Journal of Clinical Oncology Conference: 8th Asia Pacific Oncology Summit, APOS 2010 Tokyo Japan. Conference Start: 20100320 Conference End: 20100321. Conference Publication: (var.pagings). 41 (3): i12.

Duplicate
Joensuu, H., M. Eriksson, et al. (2011). "Twelve versus 36 months of adjuvant imatinib (IM) as treatment of operable GIST with a high risk of recurrence: Final results of a randomized trial (SSGXVIII/AIO)." Journal of Clinical Oncology. Conference: ASCO Annual Meeting 29(18 SUPPL. 1).

Kang, B., J. Lee, et al. (2009). "A phase II study of imatinib mesylate as adjuvant treatment for curatively resected high-risk localized gastrointestinal stromal tumors." Journal of Clinical Oncology Conference:

2009 Annual Meeting of the American Society of Clinical Oncology, ASCO Orlando, FL United States. Conference Start: 20090529 Conference End: 20090602. Conference Publication: (var.pagings). 27 (15 SUPPL. 1): e21515.

Foreign language

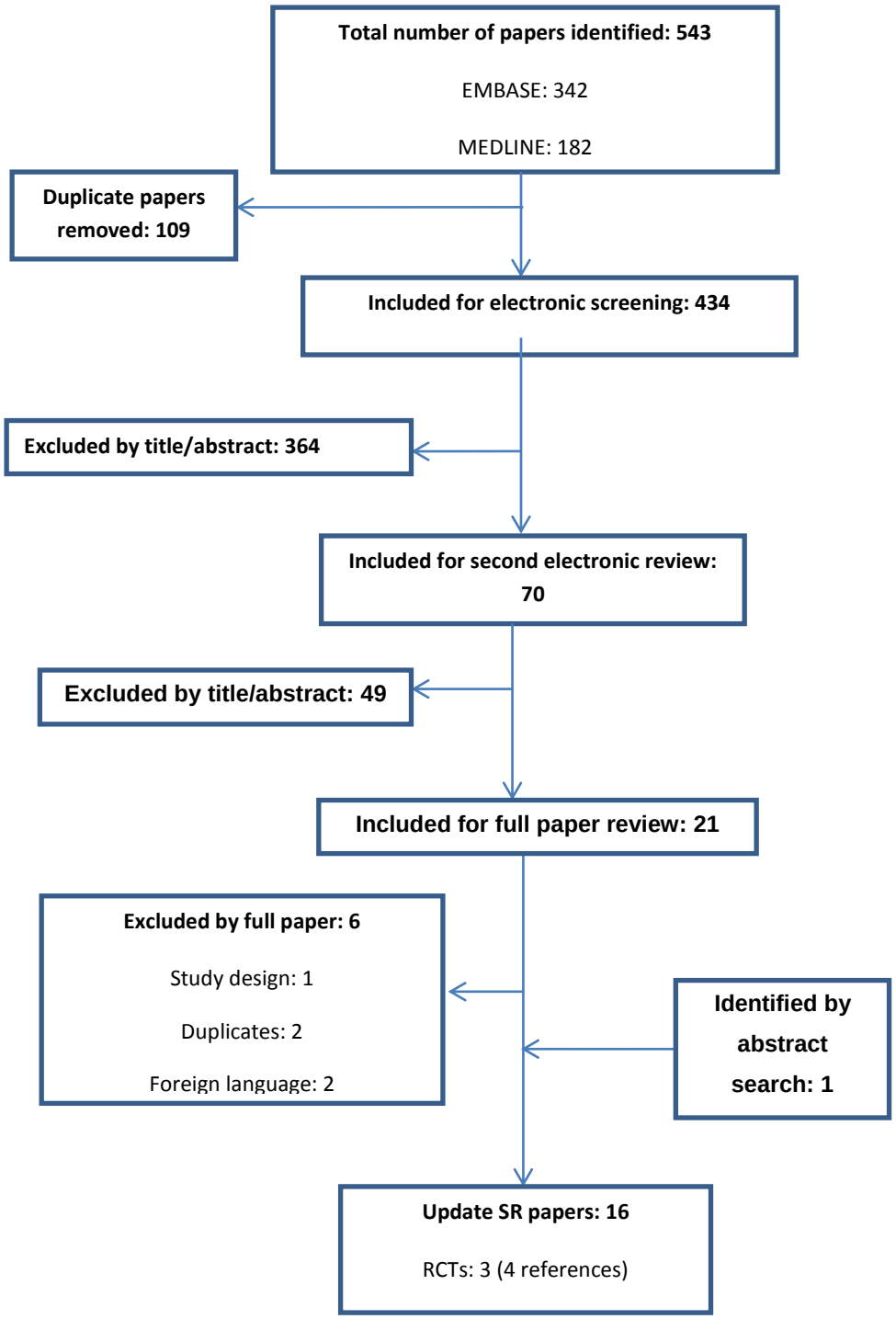
Xia, Z. F., S. F. Chen, et al. (2010). "Diagnosis, treatment and prognosis of gastrointestinal stromal tumors: An analysis of 67 cases. [Chinese]." World Chinese Journal of Digestology 18 (21): 2266-2270.

Zheng, C. H., Y. S. Zhu, et al. (2011). "Assessment of recurrence risk in gastrointestinal stromal tumors. [Chinese]." Tumor 31 (6): 549-552.

Patient population (n <20)

Vendettuoli, M., D. Pironi, et al. (2012). "Gastrointestinal stromal tumors treatment in the Imatinib era. The role of fair indication." Minerva Chirurgica 67 (2): 165-173.

Figure B-1. PRISMA diagram of included and excluded studies in the systematic review



PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses; RCT, randomised controlled trial; SR, systematic review

A5. The data for EORTC 62024 trial is only currently available as a conference abstract, which doesn't detail the baseline characteristics. Please clarify the source for the information in Table B-7.

The baseline/demographic information in Table B7 is from the slide presentation by Casali et al - specifically from slides 14 and 15 from the deck 'Casali, EORTC 62024 ASCO 2013.pptx' .

A6. Please confirm whether EORTC 62024 received funding from Novartis

In accordance with the latest contract available to the UK, Novartis have had the following commitments:

- Novartis shall provide EORTC imatinib as an adjuvant treatment to each participating centre free of charge for the specific use foreseen in the protocol for the entire study.
- Novartis shall financially support the EORTC with the maximum amount of EURO 3,004,244 for the conduct of the study.
- Reasonable travel and out-of-pocket expenses incurred by EORTC Data Center shall be covered by Novartis up to a maximum of 13,000 Euros based on mutual agreement and detailed reimbursement requests from the EORTC Data Centers.
- Any additional requests/task may be made only by prior written agreement between EORTC and Novartis.

A7. Please specify whether there are any data in the submission to inform sub-group analysis by tumour genetic mutation site (other than the data reported in the SSG XVIII journal publication – Joensuu et al 2012, JAMA, 307, 12, Figure 3).

The submission dossier does not include any further data to inform sub-group analysis by tumour genetic mutation site. However, the CSR for the 5-year follow up of the ACOSOG study has presented limited data, analysed by mutation site (see section 11.3.3.4 of the CSR, provided in our original reference pack). As can be seen, only the Exon 11 sub-group is sufficiently large to warrant separate presentation. It should also be noted that these are the results for the *overall* ITT population, not those for the Miettinen high-risk population which are the subject of this submission. Clearly, the numbers of patients for each mutation type would be further reduced if these data were considered within the high-risk sub-group.

We would also reiterate that, at present, it is not clear whether mutation type is *independent* of the risk factors that are the basis of the recommended stratification scheme (i.e., Miettinen 2006 criteria of tumour size, location and mitotic count). There is also no evidence of their use in informing the decision to use adjuvant therapy.

Section B: Clarification on cost-effectiveness data

B1. **Priority question:** The costs for patients undergoing further surgery following first recurrence are included in the model, however the outcomes for these patients are not. Please can you provide a rationale for including these costs without the outcomes (MS section 7.5.6)

Response:

The primary treatment of patients with metastatic or recurrent gastrointestinal stromal tumours (GIST) involves medical therapy with a tyrosine kinase inhibitor (TKI); in select cases, surgical therapy may be added.

Surgery is typically undertaken where there is limited disease progression (i.e., isolated tumour deposits that have progressing on TKI therapy, while other sites of disease remain stable) or where there is an immediate clinical need, such as to relieve bowel obstruction or stop bleeding. Surgery should be considered in patients with localised progression (i.e. <3 sites) to alleviate symptoms (Reid et al 2009 Scottish guidelines)

UK guidelines indicated that although surgery may be considered in patients initially considered to be inoperable, but where imatinib has led to a reduction in disease bulk such that an R0 resection may be achieved, the role of surgery in advanced disease following treatment is as yet unproven; there is no evidence at present that such surgery results in longer survival. (Reid et al UK guidelines)

B2. **Priority question:** Please confirm whether the group of patients undergoing further surgery following first recurrence are included in the *"Post recurrence and on imatinib 400mg"* health state (also labelled C). And whether they incur the costs of receiving imatinib (plus costs of monitoring patients on imatinib) and surgery, and experience outcomes (in terms of adverse events, probability of recurrence and quality of life) estimated for imatinib-treated patients. Please provide a rationale for adopting this approach

Response:

Patients undergoing further surgery are included in the *"Post recurrence and on imatinib 400mg"* health state (C). These patients incur the costs of receiving imatinib (plus costs of monitoring patients on imatinib) and surgery, and experience outcomes (in terms of adverse events, probability of recurrence and quality of life) estimated for imatinib-treated patients.

Please note, the ERG questions did not contain a *B3*. We have retained the original numbering to avoid confusion.

B4. **Priority question:** Please confirm how primary and secondary imatinib resistance is defined in the submission. Please clarify how the economic model takes imatinib resistance into account.

Response:

Primary resistance is deemed to occur when a patient does not respond to a drug on first exposure. Secondary resistance occurs when a previously responding patient goes on to lose that response. In the adjuvant setting, the model already accepts that some patients will recur whilst on adjuvant therapy, representing the rates seen in the trials. It is difficult to say whether these recurrences are due to primary resistance or secondary resistance since, in the adjuvant setting, the patient has no obvious disease to respond, not respond, or lose response. However, the response rates from the trials implicitly include these situations and so it is not necessary to take account of the resistance potential in any additional way.

Similarly, both primary and secondary resistance can occur in the metastatic setting although here, of course, one would identify resistance by progression rather than recurrence. Once

again, the model reflects the presence of resistance by using the rates of progression seen in the trials of imatinib in the advanced/metastatic setting.

The evidence that has been collected to date does not support the hypothesis that using imatinib in the adjuvant setting makes someone more likely to present with primary resistance in the advanced/metastatic setting. Nor does is suggest that someone is more likely to develop secondary resistance, or develop it sooner. Therefore, the assumptions used in the model reflect the known data on imatinib.

B5. **Priority question:** The ERG have been unable to replicate the adjusted indirect comparison presented on page 151 to 152 of the MS, to determine the off-treatment hazard ratio for recurrence. Full results including estimated standard errors (calculated by ERG) are in **Table 1**, but contrast with the value of 0.344 (0.160 to 0.741) presented in the MS (page 152). The details of the calculations are below. Please could you identify the cause of this discrepancy?

- Page 151 of the MS states that treatment effect of 3 years of adjuvant imatinib vs surgery alone (HR3) can be derived as:
  - log(HR3)=log(HR1)-log(HR2)
    - where HR1 = 1 year of adjuvant imatinib vs surgery alone
    - HR2 = 1 year of adjuvant imatinib vs 3 years of adjuvant imatinib
    - log() indicates the natural logarithm
- The final paragraph provides values, with 95% CIs, for the two hazard ratios on the right hand side of the above equation:
  - HR1 = 0.519 (0.297 to 0.906)
  - HR2 = 0.663 (0.392 to 1.123)
- Putting these values into the above equation gives:
  - log(HR3)=log(0.519)-log(0.663)= -0.656 – (-0.411)= -0.245
    - if log(HR3)= -0.245 then HR3=exp(-0.245)=0.783

Table 1 Hazard ratio for recurrence (estimated by the ERG)

	HR	**selogHR**	**95% CI**
HR1	0.519	0.28480	(0.297 to 0.907)
HR2	0.663	0.26850	(0.392 to 1.123)
HR3	0.783	0.39141	(0.363 to 1.686)

Response:

This was a documentation error, the analysis is logically and numerically correct. See below.

Taken from the submission document:

Log(HR[adj3 vs. noT]) = log(HR1) – log(HR2)

Derived from poster presented by Trowman et al
(Rebecca Trowman, Ros Collins, Rob Riemsma, Stephen Palmer: Implementing NICE guidance for evidence synthesis in technology appraisals: An overview of common methodological and analytic challenges
https://www.york.ac.uk/inst/crd/Posters/Implementing%20NICE%20guidance%20for%20evidence%20synthesis%20in%20technology.pdf)

Log(HR[D+P vs. P]) = log(HR[D+P vs. M+P]) – log(HR[P vs. M+P])

Adapted accordingly:
Log(HR[ad3 vs. noT]) =log(HR[adj3 vs. adj1]) – log(HR[placebo vs. adj1])

The underlined term is not the same as HR2, as the reference category has to be changed to adjuvant imatinib 1 yr (eg. taking the inverse of 0.519). The calculations are as follows:

Log(HR[adj3 vs. noT]) = log(0.663) – log(1.927)
Log(HR[adj3 vs. noT]) = -1.067
HR[adj3 vs. noT] = 0.344

B6. On page 130, the final two sentences of the top paragraph appear to be duplicated. Please confirm whether these are duplications or whether they refer to different groups of patients eligible for sunitinib treatment following recurrence

- the sentences start with *"The majority of patients (90%) who experience a first recurrence …"*

Response:

We confirm that the sentence was duplicated.

B7. Please clarify the number of patients whose data was analysed to estimate the baseline risk of recurrence in the comparator arm of the model.

Response:

Ninety-eight (98) patients were included in the analysis to estimate the baseline risk of recurrence on the comparator (placebo) arm. The efficacy data used was from the primary analysis (median follow-up of 19.7 months), but the additional 17 high risk patients identified at the time of the 5-year update were also included in the analysis. See also response to B8 for further details

B8. There are discrepancies in the number of patients classified as having high risk disease.

- On page 140 the MS states *"165 patients were classified as having Miettinen high-risk disease…"*
  - This agrees with Table B-13 (page 67 of MS) which indicates that, for the primary analysis, 84 imatinib-treated and 81 placebo patients were classified as Miettinen high-risk (total 165).
- Page 140 also states that *"… retrospective data were available for a further 71 patients … (of which 36 were classified as high risk), leading to … 98 patients in the placebo group."* (an additional 17 patients)
  - This figure agrees with Table B-14 (page 69 of MS) which indicates that, at five year follow-up, 103 imatinib-treated and 98 placebo patients were classified as Miettinen high-risk.

Can you confirm whether the additional 17 placebo group patients identified in the five year analysis as Miettinen high-risk were misclassified (i.e. as not being at high risk) in the primary analysis? Please also indicate the nature of the "retrospective data" that became available at five year follow-up that were not previously available for the primary analysis?

Response:

The ACOSOG Z9001 trial was stratified according to tumour size, the only known risk factor for recurrence at the time of the trial's design. However, because tumour specimens were also collected in the trial, this allowed for a retrospective analysis of mitotic count and an unpublished assessment of risk according to Miettinen 2006 criteria, which was used to answer questions raised by the European the regulatory authorities. In the timeframe for that EMA response, not all tumour analyses were possible. A summary of the available data at each time point is presented below.

ACOSOG trial: Type of analysis	All patients	Miettinen high-risk sub-group	Notes
All patients, ITT (CSR 2008, submission ref 55)	713	n/a	Risk stratification did not exist at the time the study was designed and was not part of the original statistical analysis
Sub-group analysis provided to EMA in January 2009 during the regulatory process. This is a new reference attached to this ERG response (Hatfield et al, see especially table 1-7), further details in reference 12 from this submission.	556	165 (84 imatinib and 81 placebo)	During the EMA evaluation, Novartis were asked to provide a break-down by risk group. This entailed going back to tumour specimens to assess the mitotic rate. This was not possible for all trial patients in the timeframe provided.
5-year follow up analysis provided to EMA as part of post-licensing commitments (CSR 2011, submission ref 13)	627	201 (103 imatinib and 98 placebo)	Since the first sub-group analysis, some of the missing data were recovered and this allowed additional patients to be identified. We have used this knowledge to increase the sample size of the high-risk sub-group.

The 'extra' 17 placebo patients are the difference between the 98 identified at 5-year follow up and the 81 patients who had been identified at the EMA sub-group stage. They were never *mis*-classified but, rather, *not* classified and therefore omitted.

B9. Please clarify the source of the HR for RFS 'on treatment', presented on page 143 of the MS.

- The MS states (page 147) that this is estimated from the 1-year imatinib and placebo arms of the ACOSOG Z90001 trial, calculated using Cox proportional hazards model truncating

the data at 12 months, reporting a value of 0.111 (95% CI 0.043 to 0.281) - see pages 143 and 148 of the MS.

- However, the HR reported for Meittinen high risk patients in the primary analysis is 0.265 (0.148 to 0.477) see Table B-13 (page 67 of the MS).

Is the difference between these two results primarily related to truncating at 12 months or due to the retrospective re-classification of patients at high risk?

Response:

The HR of 0.265 is valid for the primary analysis, for patients at high risk of recurrence (~Miettinen risk criteria). This HR was derived from the primary analysis sample of the ACOSOG trial (n=84 and n=81 on imatinib and placebo, respectively, at high risk) and it is based on the 5-year follow-up analysis.

The HR of 0.111 is valid for the treatment period (first 12 months /36 months) of patients at high risk of recurrence (~Miettinen risk criteria), and used in cost –effectiveness calculations. This HR was derived from the primary analysis sample updated with the retrospective data of 36 high risk patients (n=103 and n=98 on imatinib and on placebo, respectively), truncated at 12 months for the imatinib arm to estimate RFS compared to placebo (see more on the method in 7.3.2 of the MS).

B10. There are inconsistencies in the assumptions regarding the treatment effects and the incidence of adverse events, in relation to duration of imatinib treatment.

- Page 143 of the MS states *"treatment effect during the 'on treatment' period was assumed to the same irrespective of the planned treatment duration".*
- In contrast, different monthly probabilities of discontinuation are applied for patients receiving 1-year adjuvant imatinib and those having 3-year adjuvant imatinib.
- The application of different monthly probabilities of discontinuation depending on duration of planned treatment also seems to contradict the statement made in the MS (page 177) that *"[N]o notable increase in adverse events during 3 years of treatment was observed in the SSGXVII/AIO trial when compared with 1 year of treatment … suggesting that longer treatment durations are not associated with a greater incidence or severity of adverse events".*

Please provide a rationale for applying different assumptions to efficacy and adverse events.

Response:

Different monthly probabilities of discontinuation are applied for patients receiving 1-year adjuvant imatinib and those having 3-year adjuvant imatinib. We acknowledge that this is not consistent.

We have therefore provided a scenario in which we used the 1-year adjuvant imatinib discontinuation probabilities in both arms.

For the revised scenario the 6-month and 12-month cumulative incidence of discontinuation due to adverse events or death of the 1 year arm was used to estimate discontinuation rate on the 3 year adjuvant imatinib arm. Data was derived from Table 12-4, page 102 of SSG trial CSR. Discontinuation due to adverse event was assumed to remain unchanged after 6 months for the complete treatment period.

Input parameters used are shown below:

	Current basecase	Revised
6 months (1 year arm)	0.014	0.014
7-12 months (1 year arm)	0.001	0.001
6 months (3 year arm)	0.010	0.014
7-36 months (3 year arm)	0.004	0.001

The impact on the ICER was limited (See below)

	ICERs	
	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£3 509**	**£2 984**
Adj Imat 3yrs vs Adj Imat 1yr	**£16 468**	**£13 448**
Adj Imat 3yrs vs Placebo	**£8 568**	**£7 170**

An alternative scenario was also explored. As above for the first year of treatment (using 6-month and 12-month cumulative incidence of discontinuation due to adverse events or death from the 1 year arm to estimate discontinuation rate on the 3 year adjuvant imatinib arm for the first year only) but then using the discontinuation rate derived from the adjuvant 3 year treatment data applied for years 2 and 3 in the three year arm.

	Current basecase	Alternative scenario
6 months (1 year arm)	0.014	0.014
7-12 months (1 year arm)	0.001	0.001
6 months (3 year arm)	0.010	0.014
7-12 months (3 year arm)	0.004	0.001
13-36 months (3 Year arm)	0.004	0.004

Results for alternative scenario are shown below:

	ICERs	
	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£3 509**	**£2 984**
Adj Imat 3yrs vs Adj Imat 1yr	**£15 799**	**£12 902**
Adj Imat 3yrs vs Placebo	**£8 311**	**£6 955**

Once again, the impact on the ICER is limited.

B11. Adverse event probabilities for both 1-year adjuvant imatinib and 3-year adjuvant imatinib are based on the SSGXVIII/AIO study. However Table B-18 (page 102 of MS) indicates there are also adverse event data available for the ACOSOG Z9001 study (although this table presents

data for the whole population and not Miettinen high-risk sub-group). Please provide a rationale for why only the SSGXVIII/AIO study was used.

Response:

It would be justifiable to use the ACOSOG adverse event profile as an alternative to the SSG adverse event profile. However, the ACOSOG data presents relevant information only for the 1 year treatment scenario, while the use of the SSG trial allows the use of combined information from both arms. Further, the definitions of adverse events between the two trials were not consistent. For simplicity it was decided to use the SSG trial results.

The impact on the ICER of using different assumptions is expected to be minimal. In the MS, sensitivity analysis was conducted on doubled/5x AEs (SA 39 and 40). This showed that the effect of varying this assumption had no significant impact on the ICER (see pg. 224 of the MS).

B.12 Please confirm whether the hazard ratios derived for on-treatment and off-treatment effects of adjuvant imatinib were estimated using Cox proportional hazards models (page 147 of MS). And whether the hazard ratios derived using the Cox models (and further included in adjusted indirect comparisons) were applied to a variety of baseline survival functions (for surgery alone) derived using fully parametric models.

- If so, please could you provide a rationale for using hazards ratios derived using the Cox models rather than estimates derived using the relevant parametric models?

Response

The hazard ratios (for the on and off-treatment) were calculated using Cox proportional hazard models to allow a comparison between 3 years adjuvant imatinib and placebo.

Using curves from different trials is likely to introduce biases.

In the absence of head to head comparison between 3 years adjuvant imatinib and placebo an indirect comparison was undertaken requiring the treatment effect for 1 year vs placebo and 1 year vs. 3 years adjuvant imatinib in the form of a hazard ratio.

Similarly, the use of a hazard ratio for the on-treatment period was necessary to ensure that the hazard of recurrence would be the same at year 1 for patients receiving 1 year and 3 years adjuvant imatinib. Using the curves from the trial would lead to inconsistencies.

The calculated hazard ratios were applied to a baseline parametric survival curve (placebo) to estimate the hazard of recurrence for patients receiving 1 and 3 years adjuvant imatinib. Different parametric extrapolation (exponential, weibull, gompertz) were tested in sensitivity analysis. (MS p 141)

B.13 Page 155 of the MS states that the monthly probability of discontinuing adjuvant imatinib *"was assumed to be different in the first 6 months of adjuvant treatment and next 6 months".* Please could you provide a rationale for choosing those particular time points?

Response

In the base-case the model uses the 6-months' and 12-months' cumulative incidence of discontinuation due to AE or death on the one-year adjuvant imatinib arm, and equivalent data at 6 months and 36 months on the three-year arm data.

The selection of the 6 month cut-off was data driven, although this was limited to the 1 v 0 comparison from the ACOSOG study. In the context of SSG study this trend is not seen to the same extent and so the rationale for the 6 month cut-off is not so clear cut. We have, therefore, presented an alternative analysis.

A sensitivity analysis using all the data points presented (See table12-4 below) shows that the ICERs are reduced.

Table 12-4 Cumulative percentage estimates of patients discontinuing early due to adverse event or death (safety population)

	Imatinib	
	12 months N=194	36 months N=198
Cumulative %*	% (95% CI)	% (95% CI)
By 4 months	5.8 (3.2 - 10.2)	4.1 (2.1 – 8.0)
By 6 months	7.9 (4.8 – 12.8)	5.7 (3.2 – 10.0)
By 8 months	8.4 (5.3 – 13.4)	5.7 (3.2 – 10.0)
By 10 months	8.4 (5.3 – 13.4)	5.7 (3.2 – 10.0)
By 12 months	8.4 (5.3 – 13.4)	7.3 (4.4 – 12.0)
By 18 months	-	9.0 (5.7 – 14.1)
By 24 months	-	9.6 (6.1 – 14.7)
By 30 months	-	10.9 (7.1 – 16.4)
By 36 months	-	16.6 (11.7 – 23.2)

*The cumulative percentage estimates relate to the probability of discontinuing treatment early due to AE or death by the specified timepoint

Source: Table 14.3-1.9

	ICERs	
	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£3 493**	**£2 970**
Adj Imat 3yrs vs Adj Imat 1yr	**£13 862**	**£11 320**
Adj Imat 3yrs vs Placebo	**£7 554**	**£6 322**

B.14 Page 155 of the MS presents the proportion of patients discontinuing adjuvant imatinib treatment, but does not indicate the reason for discontinuation. On a previous page (page 153, Table B-29) the monthly probabilities derived from these discontinuation values are stated as being 'discontinuation due to adverse events'. Please clarify whether these are all discontinuations or only discontinuations due to adverse events.

Response:

These data are exactly the same (the values appearing on p153 of MS are transitional probabilities derived from values appearing on p155), they originated from Table 12-4, page 102 of SSG CSR. The correct definition for them is "discontinuation due to adverse event or death".

NOTE: This is not double counting because the disease-specific mortality is set to zero while on adjuvant imatinib treatment.

B.15 Please clarify what is meant by *"Recurrence was used as a surrogate outcome for death"* (see first paragraph under section 7.3.4; top of page 157)

Response:

The aim of adjuvant imatinib is to prevent recurrences. Recurrence free survival (RFS) is the primary outcome in the two key trials - AGOSOG and SSG. The model predicts the long term outcome for RFS, based on the trial results reported to date. Once patients have experienced a recurrence the treatment pathway followed is assumed to be the same, irrespective of the initial therapy. Therefore recurrence, or the avoidance of it, is the key determinant of overall survival.

B.16 Treatment effectiveness is assumed to continue in the off treatment period, please provide a justification for this assumption.

- Please provide scenario analyses that assumes treatment effect declines over time for the off treatment period.

Response:

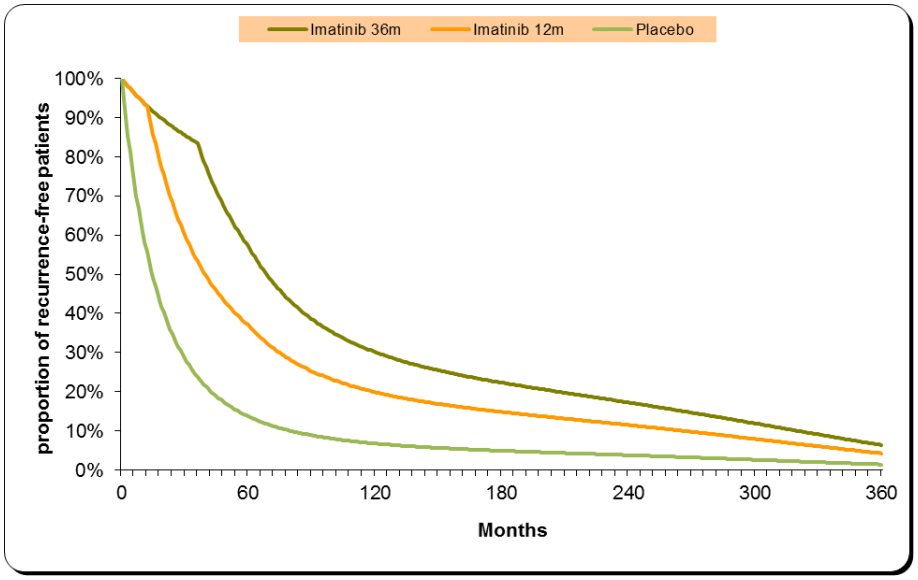
The treatment effect for the "off treatment period" reflects the period after cessation of adjuvant imatinib and is based on the clinical data from the pivotal phase III trials. Results for the SSG are reported for a median duration of follow-up of 54 months and therefore offer robust data for 5 to 6 years. These demonstrate that that the hazard of recurrence in the imatinib arm is below that seen in the placebo arm for the off treatment period.

As requested, the following scenarios demonstrate the impact of assuming that the treatment effect declines over time for the off treatment period.

Scenario A : Off treatment HR reduced to 75% after 5 years:

	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£4 569**	**£3 880**
Adj Imat 3yrs vs Adj Imat 1yr	**£18 242**	**£14 818**
Adj Imat 3yrs vs Placebo	**£9 952**	**£8 302**

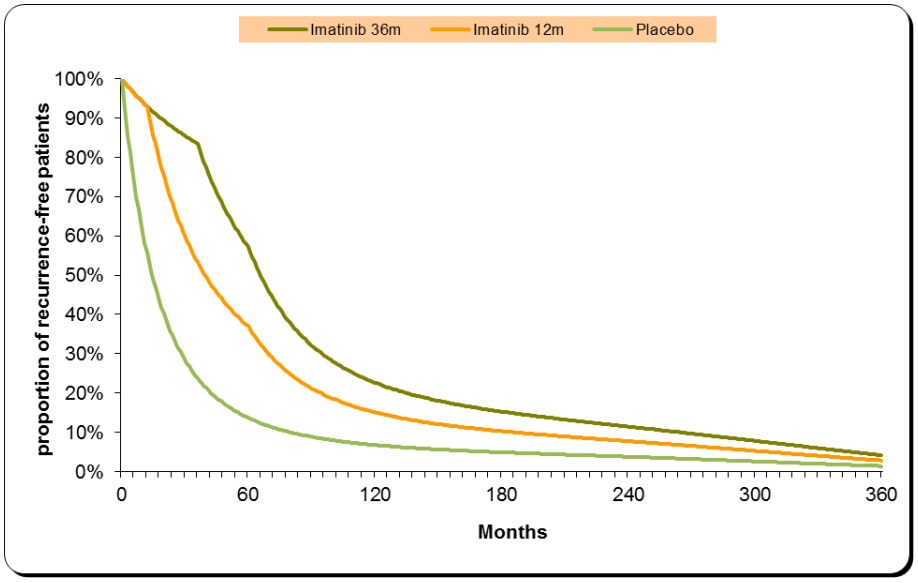
Predicted RFS (Scenario A)



Scenario B : Off treatment HR halved after 5 years:

	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£6 831**	**£5 783**
Adj Imat 3yrs vs Adj Imat 1yr	**£22 735**	**£18 277**
Adj Imat 3yrs vs Placebo	**£13 210**	**£10 951**

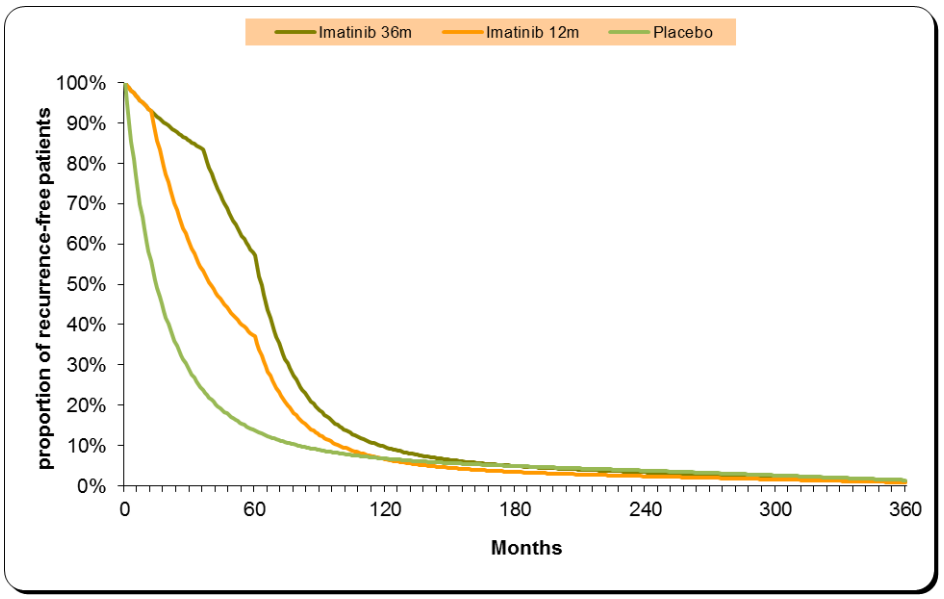
Predicted RFS (Scenario B)



Scenario C : Off treatment HR reduced to 25% after 5 years:

	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£14 079**	**£11 821**
Adj Imat 3yrs vs Adj Imat 1yr	**£34 683**	**£27 153**
Adj Imat 3yrs vs Placebo	**£22 939**	**£18 678**

Predicted RFS (Scenario C)



B.18 The model assumes a long-term survival benefit for imatinib, and that the post-recurrence GIST mortality is equal across all treatment arms. Please provide justification for these assumptions.

- Please provide scenario analyses in which the post-recurrence GIST mortality is different between treatment arms, such that the mortality probability is higher on the adjuvant imatinib arms than the control arm.
- Following further explanation on B18 by the ERG, please could you repeat the sensitivity analysis carried out in your previous submission (that is, the submission for TA196). This is described in the West Midlands ERG report on page 69: 'the monthly probability of death in the recurrent state was changed so as to be greater in the adjuvant arm than the control arm.' In the current model this is state C (post recurrence and on imatinib 400mg). (text copied form the email of 11th March 2014 to Novartis from Nicole Fisher)

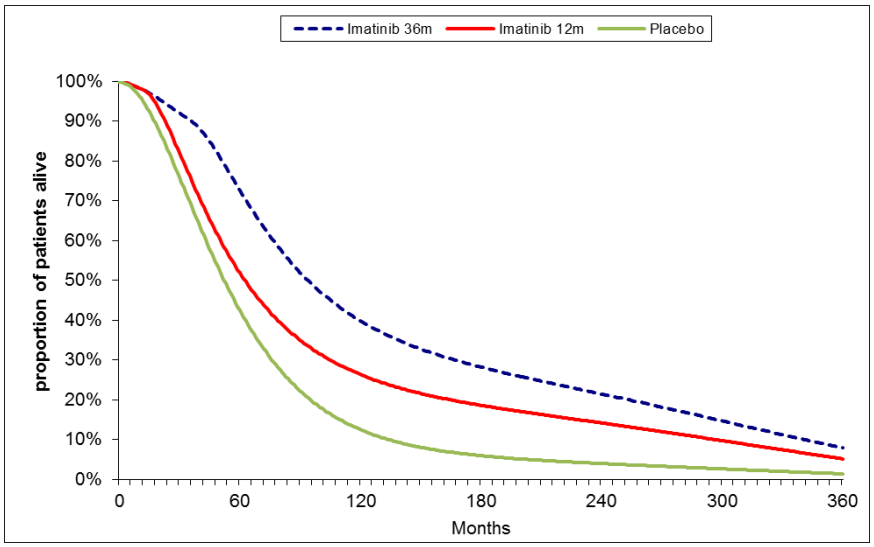
Response

In the base case, it is assumed that the benefits of imatinib in the metastatic setting are the same whether or not patients have had imatinib in the adjuvant setting. To date there is no evidence to suggest that this is not case.

The SA referred to in the second bullet above is SA 15 from our first submission for TA196. In this SA, the mortality was increased 4-fold in the adjuvant arm for patients in the 'post recurrence and on imatinib 400 mg' health state – Health State C. Results are presented below:

	Total Per Patient:				Incremental:		
	Costs	QALYs	Life-Years		Costs	QALYs	Life-Years
No treatment	£47 292	3.83	5.08				
Adjuvant imatinib 1 yr tx	£41 829	5.55	6.95		-£5 463	1.72	1.87
Adjuvant imatinib 3 yr tx	£69 123	7.15	8.95		£27 294	1.60	2.00

	ICERs	
	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**Dominant**	**Dominant**
Adj Imat 3yrs vs Adj Imat 1yr	**£17 036**	**£13 624**
Adj Imat 3yrs vs Placebo	**£6 573**	**£5 634**



In this scenario, the LYs gained in the 1 year and 3 year arms are reduced from 7.71 and 9.46 to 6.95 and 8.95, respectively. In the one year arm the costs savings from avoided costs on treatment in the metastatic setting (imatinib, sunitinib and BSC) reduces the overall costs to be below those in the placebo arm so the 1 year treatment arm becomes dominant. In the three year arm the higher costs in the adjuvant setting mean the impact on overall costs is reduced.

Section C: Clarification on Process and Confidentiality

C1. **Priority question:** It has been noted that your submission refers to an analysis of crossover data which is currently on-going. It is indicated that results and analyses will be updated and

these data incorporated. In line with NICE process, to submit further evidence you will need to make a request to do so, and this is subject to agreement from NICE. To clarify, there may be an opportunity to submit further information at the ACD consultation comments stage: the NICE guide to the STA process published October 2009, states in section 3.5.34: ' *At the ACD consultation stage the Centre Director must agree to accept any new evidence before it is submitted. New evidence will only be accepted if it is likely to affect the provisional recommendations in the ACD. ….*'. Please clarify the timings that both the data and analyses will be available. If this data or analysis is available now, please provide it.

Response:

We confirm that the cross-over analysis report is complete and that, following a formal, written request, NICE have granted permission for it to be submitted for the ERG's review (*see Appendix A to this document*). We are unable to use the results from the report to undertake any additional modelling since this is not within the process. Any such analysis will now be allowed only at ACD stage.

In the absence of being able to investigate the effect of the cross-over on the cost-effectiveness results, we have instead reproduced the base-case analysis using the 5-year update data from the ACOSOG study, *without* accounting for the cross-over. The HR for RFS is 0.61 (whereas the adjusted HR, taken from the cross-over analysis report, is 0.5).

The baseline placebo and 1 year KM curves (and HRs) are taken from the clinical trial data (Figure 11-5 on p70 of ACOSOGZ9001/STI571BUS89 CSR). The on-treatment and off-treatment HRs are derived and applied in the model using the same methodology (and, for completeness, the indirect comparison HR is also recalculated and applied).

	Current basecase	5 year update Unadjusted
HR on treatment	0.111	0.112
HR off treatment (1 year imat)	0.519	0.727
HR off treatment (3 year imat)	0.344	0.482

The ICER based on the 5 year update (unadjusted for crossover) is as follows:

	ICERs	
	Cost/QALY Gained	Cost/LY Saved
Adjuvant Imat vs Placebo	**£8 556**	**£7 274**
Adj Imat 3yrs vs Adj Imat 1yr	**£17 057**	**£13 930**
Adj Imat 3yrs vs Placebo	**£12 975**	**£10 800**

C2. **Priority question:** Much of the data that will be critical to the decision-making is currently marked as academic in confidence. The NICE STA process guide states that confidential information should be kept to an absolute minimum to ensure transparency in the appraisal process. As per compliance with CONSORT, that is, that as a minimum

a structured abstract should be made available for public disclosure (please refer here http://www.consort-statement.org/?o=1011 for further information), further details need to be made publically available. Examples of such data include the arms of the study and number of people on each arm (currently marked as confidential on page 57 of the submission); particularly given these data are currently available in the public domain. Please reconsider the marking in your submission to ensure that:

- no data that are available in the public domain are marked as confidential
- a structured abstract (for all trials presented) is available for public disclosure

An updated submission and redacted version is provided, lifting the restriction on page 57 and addressing those inconsistencies which have been identified.

As previously communicated, we are working with the legal department in the Novartis Global organisation to understand the terms of the contracts which govern the rights to the data for the ACOSOG and SSG studies. These terms will need to be re-negotiated with the ACOSOG and SSG organisations in order for the data currently marked as AIC to be made public. We cannot guarantee the availability of the data by a particular date since such negotiations may be protracted; moreover, we cannot insist that the restriction is lifted.

C3. **Priority question:** The confidential marking in the submission is currently inconsistent, the same information is marked as confidential in one area but not in another. NICE takes the protection of your confidential information seriously and discrepancies can present a risk of accidental disclosure. Please be consistent and clear in your marking, and limit marking only to items that are confidential (for example, items such as figure titles should not be confidential). Please readdress the marking to ensure that:

- the marking is consistent
- the marking is appropriate

An updated submission and redacted version is provided, addressing those inconsistencies which have been identified.

Appendix A – additional response to C1

Adjusting for treatment switching in a Phase 3 trial of imatinib for the adjuvant treatment of adult patients who are at high risk of relapse following resection of kit (CD117)-positive gastrointestinal stromal tumours (GIST)

Prepared by: Nicholas Latimer and Sue Ward

School of Health and Related Research (ScHARR), University of Sheffield

Contact details:

Nicholas Latimer

School of Health and Related Research (ScHARR)

University of Sheffield

[redacted]

Phone: [redacted]

Email: [redacted]

Date: 7th March 2014

Executive Summary

In this report we present analyses that attempt to adjust for the treatment switching observed in the Phase III trial of imatinib for the adjuvant treatment of adult patients who are at high risk of relapse following resection of kit (CD117)-positive gastrointestinal stromal tumours (GIST). First we introduce the problems caused by treatment switching, and describe the methods available to adjust for treatment switches, before focussing upon the clinical trials in question.

Following resection of primary GIST, patients were randomised to one of two arms: imatinib at 400 mg/day or matching placebo for one year. Upon disease recurrence, patients were to be unblinded: In the imatinib arm, if the disease recurred during the year of initial treatment, the imatinib dose was to be increased to 800 mg/day. If the disease recurred after the year of initial treatment, the drug was to be restarted at 400 mg/day and could be increased to 800 mg/day; In the placebo arm, if the disease recurred at any time, imatinib was to be started at 400 mg/day and could be increased to 800 mg/day. Hence, the trial planned for placebo group patients to receive imatinib after disease recurrence. To an extent, such post-recurrence use of imatinib represents usual care in the UK. However, the availability of sunitinib means that the sequences of treatments received in the imatinib trial may not be precisely what would be expected presently. This is because imatinib was a new treatment at the time of the trial set-up, and sunitinib was not available. Inherently, this makes the overall survival data collected within the trial of questionable relevance, because survival could be impacted upon by different post-recurrence sequences of treatments. However, if we assume that similar treatment sequences were given in the imatinib trial post-recurrence, and that these have similar effects in each treatment group, the trial may still furnish us with appropriate estimates of the overall survival benefits associated with a treatment sequence beginning with imatinib compared to one beginning with placebo. Given this, we do not wish to adjust for the imatinib treatment received in the control group post-disease recurrence. If we did attempt to adjust for this, in the interests of consistency we would also need to adjust for post-recurrence imatinib treatment in the experimental group, but this is not possible because data on this were not collected in the trial.

However, additional pre-disease recurrence switching was permitted upon unblinding of the trial on 12th April 2007. Patients randomised to the control group who were still receiving placebo as of 1st April 2007, and those randomised to placebo between that date and 12th April 2007 were eligible to switch onto one year of imatinib treatment. Therefore, by definition, switching in these patients must have occurred prior to disease recurrence, unless recurrence occurred between 1st April 2007 and 12th April 2007. Such pre-recurrence switching confounds the comparison of imatinib compared

to placebo in the pre-recurrence period. Hence, it is relevant to adjust analyses of both recurrence free survival (RFS) and overall survival (OS) to attempt to avoid the confounding caused by control group patients who switched onto imatinib before disease recurrence.

Although the imatinib trial in was reasonably large (n=713 in the ITT population), the high-risk patient group of interest was much smaller (n=201, 98 in the control group and 103 in the imatnib group). 17 control group patients switched onto imatinib before disease recurrence, and none of those had died as of the final trial follow-up. Of all 201 patients, only 34 had died, and 110 had experienced disease recurrence. The application of switching adjustment methods becomes more problematic with low event and patient numbers, and so the results of our analysis must be interpreted with caution – particularly for the OS analysis.

Twenty control group patients were eligible to switch on to imatinib treatment, as they remained on placebo at 1st April. Of these, 18 patients switched onto imatinib. However, one of the patients who switched onto imatinib experienced disease recurrence on 23rd April and only begun imatinib treatment on 30th April, and hence actually switched after disease recurrence. Also, two of the patients who were eligible to switch but did not, experienced disease recurrence between the 2nd and the 12th April – hence if these patients *had* switched, they would have done so after disease recurrence. Therefore, if we consider only those patients who became eligible to switch onto imatinib before disease recurrence, there were 18 patients who were eligible and 17 switchers.

The inverse probability of censoring weights (IPCW) method for adjusting for treatment switching involves censoring switchers at the time of switch, and weighting remaining control group patients according to their similarity to switchers based upon baseline and time-dependent covariates. Models must be built that essentially estimate the probability of switching, but this can only usually be reliably done using data from patients who became at risk of switching. Hence, in the case of the imatinib trial, a model would be required that estimated the probability of switching in those that became eligible to switch – however, such a model would be extremely prone to error because 17 out of 18 eligible patients *did* switch. Therefore, it is not sensible to apply a standard IPCW analysis to this dataset.

However, owing to the characteristics of the imatinib trial, an alternative application of the IPCW method is possible. Because almost all the patients who became eligible to switch did so, and because the chance of switching was associated only with whoever remained on placebo treatment at the time of unblinding of the study, it could be argued that the switching was essentially random, because time of enrollment and randomisation is random. Hence, an IPCW analysis that applied

weighting models to *all* control group patients, including covariates for time since randomisation, and baseline characteristics (to correct for any imbalances in these), and incorporating the requirement that switching could only occur before disease recurrence, may represent a reasonably straightforward answer to the switching problem in the imatinib trial. In fact, if switching was truly random, then simple per-protocol approaches such as censoring patients at the time of switch, or excluding their data altogether, may also provide unbiased results. However, the IPCW may be preferable as it allows us to adjust for any potential chance imbalances in the baseline characteristics of those control group patients who switched compared to those who did not switch, and also accounts for the time since randomisation at which each patient switched.

Applying randomisation-based switching adjustment methods – such as the rank preserving structural failure time model (RPSFTM) and the iterative parameter estimation (IPE) algorithm – to the imatinib trial dataset is problematic due to the treatment pathways observed in the trial. A standard RPSFTM or IPE analysis involves assuming that at all time points throughout the trial patients are either "on" treatment or "off" treatment. The models cannot distinguish between being "on" treatment A, being "on" treatment B, or being "off" treatment altogether. This is problematic in the imatinib trial because although patients in the control group begin on placebo, after disease recurrence most patients move onto imatinib treatment, and patients in both groups received an increased dose of imatinib if required. If the different doses of imatinib can be expected to have different treatment effects, and if receiving imatinib after disease recurrence is expected to have a different treatment effect to receiving it before recurrence, then the standard RPSFTM/IPE models will not be appropriate. However, similarly, receiving imatinib treatment after disease recurrence cannot be considered being "off" treatment. We have applied variations of the RPSFTM and IPE methods under additional restrictive assumptions, described in Section 2. Although these have important limitations, we believe that the RPSFTM "treatment group (i)" analysis without recensoring provides plausible estimates of the treatment effect of imatinib on RFS, and the RPSFTM "treatment group (ii)" analysis provides plausible estimates of the treatment effect of imatinib on OS. Analyses presented in Section 4 of this report demonstrate that there is not strong evidence against the key assumptions associated with these analyses.

It is important to note that all adjustment analyses are likely to provide only approximations of the "true" treatment effect. We conclude that the most reliable switching adjustment analysis is provided by the IPCW method both for RFS and OS. For RFS the IPCW hazard ratio (HR) is 0.50 (95% confidence intervals (CI) 0.32 – 0.78). This is similar to the simple, unweighted per-protocol censoring approach, which gives a HR of 0.52 (95% CI 0.35-0.77). Both provide a numerically lower

HR than the intention to treat (ITT) analysis (HR 0.61, 95% CI 0.42 – 0.89). We believe the RPSFTM is less suited to this case study, but the RPSFTM "treatment group (i)" analysis (without recensoring), which we believe represents the most appropriate randomisation-based adjustment analysis of the imatinib RFS dataset, provides a HR of 0.52 (95% CI 0.32 – 0.85). For OS the IPCW HR is 0.76 (95% CI 0.36 – 1.62), which is again similar (but marginally lower) than the simple, unweighted per-protocol censoring approach, which gives a HR of 0.79 (95% CI 0.40-1.55). Both provide a numerically lower HR than the intention to treat (ITT) analysis HR of 0.93 (95% CI 0.47 – 1.83). Again, we believe the RPSFTM is less suited to this case study, but the RPSFTM "treatment group (ii)" analysis, which we believe represents the most appropriate randomisation-based adjustment analysis of the imatinib OS dataset, provides a HR of 0.82 (95% CI 0.42 – 1.60).

These results should be regarded as exploratory, given that they are based upon a relatively small dataset in which few events – particularly deaths – have been observed. The IPCW results are dependent upon key assumptions around how random the observed treatment switching was, and whether data on all relevant prognostic covariates were collected at baseline. These assumptions are untestable, but our analysis of baseline covariates presented in Section 4 suggests the IPCW analyses may be satisfactory, and are likely to be preferable to a simple per-protocol censoring approach. The fact that the adjusted treatment effect point estimates for RFS and OS lie relatively close together when comparing the adjustment methods that are best suited to the imatinib trial characteristics (HRs of 0.50 – 0.52 for RFS, and 0.76 – 0.82 for OS) suggests that it is reasonable to be confident in the robustness of these results. In addition, while we acknowledge the potential limitations in the switching adjustment analyses, it is also important to reiterate that we know the ITT analyses to be confounded by pre-recurrence treatment switching – hence a reliance upon this will underestimate the benefits associated with imatinib.

Contents

Page

1 Introduction **7**
2 Methods **11**
2.1 Rank Preserving Structural Failure Time Model 11
2.2 Iterative Parameter Estimation Algorithm 19
2.3 Inverse Probability of Censoring Weights 20
2.4 Two-stage methods… 24
2.5 Methods Summary 24
2.6 Methods Applied in this Report 25
3 Results **27**
3.1 Results: RFS 27
3.2 Results: OS 30
4 Key Issues **34**
4.1 Modelling approaches 34
4.2 Common treatment effect in pre-recurrence switchers 37
4.3 Common treatment effect in post-recurrence switchers 37
4.4 Baseline characteristics of switchers 38
5 Conclusions **41**

References 44

Appendix 1: Log-rank test results for alternative values of psi 46
Appendix 2: Kaplan-Meier curves from RPSFTM and IPE analyses 50
Appendix 3: Weighted Kaplan-Meier curves from IPCW analyses 59

1. Introduction

The purpose of this report is to present the results of treatment switching adjustment analyses undertaken on the Phase III trial of imatinib for the adjuvant treatment of adult patients who are at high risk of relapse following resection of kit (CD117)-positive gastrointestinal stromal tumours (GIST). We also discuss the plausibility and interpretation of the alternative switching adjustment methods and results. In this section of the report we briefly introduce the imatinib trial in question, and the treatment switching problem. In Section 2 we describe potentially appropriate treatment switching adjustment methods, and address these specifically in the context of the imatinib trial. Section 3 reports the results of our adjustment analyses, and Section 4 considers the key issues associated with the analyses undertaken. Section 5 provides a summary of the results and makes conclusions.

In the imatinib trial, following resection of primary GIST, patients were randomised to one of two arms: imatinib at 400 mg/day or matching placebo for one year. Upon disease recurrence, patients were to be unblinded: In the imatinib arm, if the disease recurred during the year of initial treatment, the imatinib dose was to be increased to 800 mg/day. If the disease recurred after the year of initial treatment, the drug was to be restarted at 400 mg/day and could be increased to 800 mg/day; In the placebo arm, if the disease recurred at any time, imatinib was to be started at 400 mg/day and could be increased to 800 mg/day. Hence, the trial planned for placebo group patients to receive imatinib after disease recurrence. To an extent, such post-recurrence use of imatinib represents usual care in the UK. However, the availability of sunitinib means that the sequences of treatments received in the imatinib trial may not be precisely what would be expected presently. This is because imatinib was a new treatment at the time of the trial set-up, and sunitinib was not available. Inherently, this makes the overall survival data collected within the trial of questionable relevance, because survival could be impacted upon by different post-recurrence sequences of treatments. However, if we assume that similar treatment sequences were given in the imatinib trial post-recurrence, and that these have similar effects in each treatment group, the trial may still furnish us with appropriate estimates of the overall survival benefits associated with a treatment sequence beginning with imatinib compared to one beginning with placebo. Given this, we do not wish to adjust for the imatinib treatment received in the control group post-disease recurrence. If we did attempt to adjust for this, in the interests of consistency we would also need to adjust for post-recurrence imatinib treatment in the experimental group, but this is not possible because data on this were not collected in the trial.

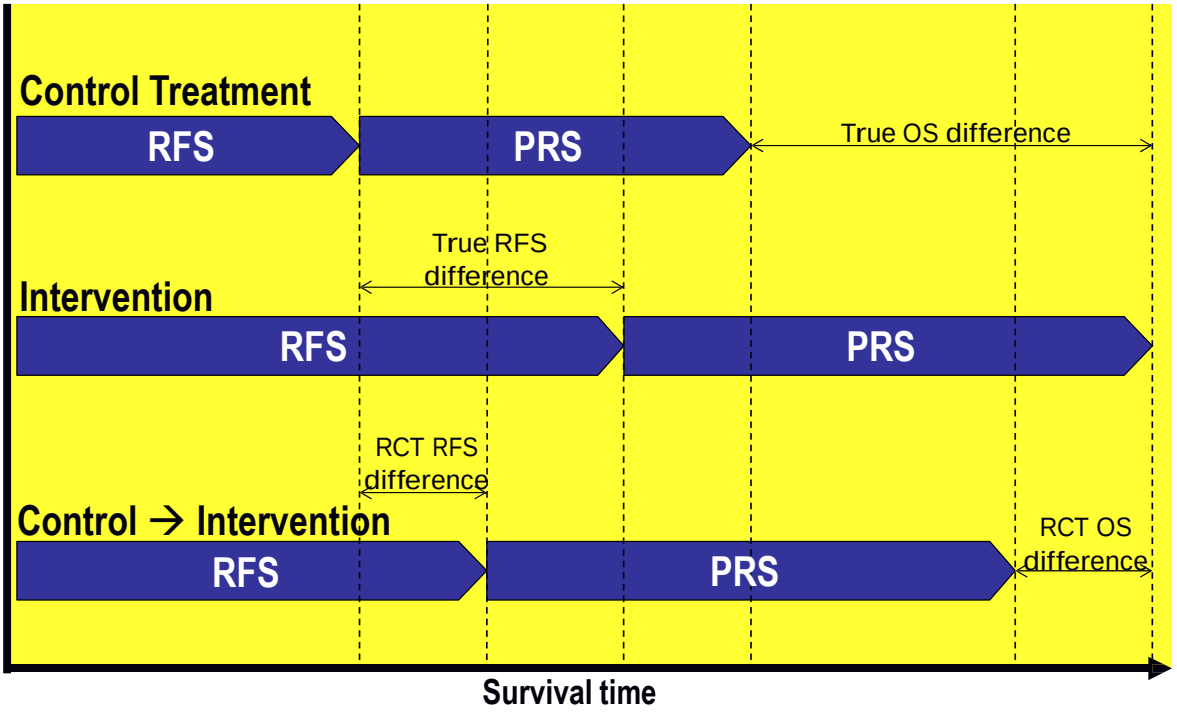
However, additional pre-disease recurrence switching was permitted upon unblinding of the trial on 12th April 2007. Patients randomised to the control group who were still receiving placebo as of 1st April 2007, and those randomised to placebo between that date and 12th April 2007 were eligible to switch onto one year of imatinib treatment. Therefore, by definition, switching in these patients must have occurred prior to disease recurrence, unless recurrence occurred between 1st April 2007 and 12th April 2007. Such pre-recurrence switching confounds the comparison of imatinib compared to placebo in the pre-recurrence period. Hence, it is relevant to adjust analyses of both recurrence free survival (RFS) and overall survival (OS) to attempt to avoid the confounding caused by control group patients who switched onto imatinib before disease recurrence.

Treatment switching is an important problem for economists and decision-makers because, unlike many post-study treatment changes, it typically does not reflect a treatment pathway that is relevant for the decision problem defined in a Health Technology Assessment (HTA).[1] Generally the decision problem requires an analysis of the impact (in terms of costs and effects) of a new treatment compared to a standard treatment. Typically randomised controlled trial (RCT) data can be used to inform these analyses, where one group of patients is randomised to receive the new treatment, and the other group is randomised to receive the standard treatment. However, when treatment switching occurs a mismatch arises between what has been studied in the clinical trial and the HTA decision problem. We wish to compare two states of the world, one in which the new treatment exists and one in which it does not. The comparator arm, which normally represents the "does not exist" state, is contaminated in the presence of treatment switching.

In the case of the imatinib trial focussed upon in this report, this distinction between the two "states of the world" is less clear cut, because the original trial protocol allowed control group patients to switch onto imatinib post disease recurrence, and indeed such treatment is common in the UK. Hence the experimental treatment exists in both "states of the world" in the HTA decision problem. However, in the control group "state of the world", imatinib is only available after disease recurrence, and any imatinib treatment received in the control group before disease recurrence contaminates this state of the world. Hence, it is specifically the treatment switching from placebo onto imatinib that occurred before disease recurrence that we wish to adjust for. Because this switching occurred before disease recurrence it is relevant to adjust analyses for both recurrence free survival (RFS) and overall survival (OS).

The bias that may be created by the treatment switching observed in the imatinib triak, and the theoretical problems that it creates for the economic analysis are illustrated in Figure 1. The first two rows ("Control Treatment" and "Intervention") illustrate the "perfect" RCT, where no treatment switching occurs. Survival time is on the x-axis, and in this example the new intervention extends recurrence free survival (RFS) and also extends post recurrence survival (PRS). This results in the "True OS difference" identified in the diagram, where OS refers to overall survival. In this case, an intention to treat (ITT) analysis will give us the information that we need for our economic model (ignoring any need for extrapolation). However, the third row ("Control → Intervention") demonstrates what may happen to survival in the control group if treatment switching is permitted in some patients before disease recurrence. RFS is extended because switching is allowed in some patients before disease recurrence. PRS is also extended compared to the "Control Treatment" comparator, under the assumption that the experimental treatment extends both RFS and PRS. The result of this is that both RFS and OS differences observed in the RCT ITT analysis (labelled "RCT RFS difference" and "RCT OS difference" in Figure 1) is smaller than the true differences that would have been observed if no treatment switching had occurred. This demonstrates that a simple ITT analysis will result in bias equal to the difference between the "True RFS difference" and the "RCT RFS difference" for the RFS analysis, and the difference between the "True OS difference" and the "RCT OS difference" for the OS analysis, when treatment switching occurs. The extent of this bias will be unknown, as the true RFS and OS differences will be unobserved. However it is clear that provided switching patients benefit to any extent from the new intervention, some bias will exist (where "bias" is the difference (error) between the ITT estimated treatment effect and the "true" treatment effect that would have been observed if there had been no treatment switching). An economic evaluation that relied upon the ITT analysis would produce inaccurate cost-effectiveness results (in this case the ICER would be over-estimated) and inappropriate resource allocation decisions may be made.

Figure 1: The potential impact of treatment switching illustrated



Notes: RFS = Recurrence Free Survival; PRS = Post Recurrence Survival; OS = Overall Survival; RCT = Randomised Controlled Trial

It is therefore important to attempt to adjust for treatment switching so that appropriate estimates of the treatment effect can be used in the economic model. However, adjusting for treatment switching is problematic. Simple approaches whereby switching patients are excluded or censored from the analysis have often been used in past appraisals undertaken by the National Institute for Health and Care Excellence (NICE),[1] but these approaches are highly prone to selection bias if treatment switching is likely to be associated with prognosis in some way. In this case excluding or censoring switchers will harm the balance between treatment groups achieved through randomisation.[1,2] Recently, more complex switching adjustment methods have been used in NICE appraisals. These include the Rank Preserving Structural Failure Time Model (RPSFTM), and Inverse Probability of Censoring Weights (IPCW).[3,4] However, these methods also have important limitations, which are briefly discussed in Section 2 of this report.

Although the imatmib trial was reasonably large (n=713 in the ITT population), the high-risk patient group of interest was much smaller (n=201). In addition, only 17 control group patients switched onto imatinib before disease recurrence, and none of those had died as of the final trial follow-up. Of all 201 patients, only 34 had died, and 110 had experienced disease recurrence. Even in the absence of treatment switching estimates of the treatment effect – particularly for OS – would be highly uncertain. In the presence of switching, uncertainty is further increased.

2. Methods

2.1 Rank Preserving Structural Failure Time Model

In 1991 Robins and Tsiatis developed the RPSFTM specifically for estimating causal effects in the presence of non-compliance in an RCT setting.[5] Their method was developed to deal with situations in which patients did not take the prescribed amount of a treatment, but can also be used in cases where patients take treatments that were not assigned to them. Thus, the method is relevant for dealing with treatment switching. The RPSFTM uses a counterfactual framework to estimate the causal effect of the treatment in question, based upon the randomisation of the trial. It is assumed that if two patients have the same observed event time and neither have received treatment, those two patients would also have the same event time if they both received treatment. This assumption is linked to the associated assumptions that the treatment effect is equal for all patients no matter when the treatment is received (the "common treatment effect" assumption), and that the randomisation of the trial means that there are no differences between the treatment groups, apart from treatment allocated.[5] The RPSFTM method has been shown to produce very low levels of bias when its assumptions are satisfied.[1,2]

The method splits the observed event time (T_i) for each patient into two, that is the event time when the patient is on the control treatment (T_{Ai}), and the event time when the patient is on the intervention treatment (T_{Bi}). For patients who are randomised to the intervention treatment, and who do not switch onto the control treatment (that is, when compliance is full in the treatment group), T_{Ai} is equal to zero. For patients randomised to the control group who do not switch onto the intervention (i.e. compliance is full in the control group) T_{Bi} is equal to zero. However, for patients who switch treatments (for whom compliance is imperfect) both T_{Ai} and T_{Bi} will be greater than zero.

The RPSFTM method relates T_i to the counterfactual event time (U_i) with the following causal model:

(1)

Here represents the acceleration factor (AF) associated with the intervention. By defining a binary process which equals 1 when a patient is on the intervention treatment, and equals zero when the patient is on control treatment, the causal model can be rewritten as:

$\int$ [ ] (2)

The value of is estimated using g-estimation. U_i is estimated using the causal model for each value of , and the true value of is that for which U( ) is independent of randomised groups. As

described by Mark and Robins (1993), a log-rank or Wilcoxon test can be used for the RPSFTM g-test in a non-parametric setting, testing the hypothesis that the baseline (untreated) survival curves are identical in the two treatment groups, or a Wald test could be used for parametric models.[6] The point estimate of is that for which the test (z) statistic equals zero.

White *et al.* (1999) demonstrate that censoring is problematic for the RPSFTM.[7] They define C_i as the administrative censoring time which corresponds to the end of follow-up and which is known for all patients, and show that the censoring time for $U_i(\ )$ is given by:

$$\int \quad [\quad] \qquad (3)$$

However, this demonstrates that depends upon X_i (treatment received), which may depend upon prognostic factors. This is because a positive or negative treatment effect may increase or decrease the probability that the survival time of an individual is censored, and, where treatment switching occurs, treatment received is likely to be associated with prognosis. In turn, this means that D_i may depend on prognostic factors, and thus the censoring of $U_i(\ )$ is informative.[7] White *et al.* suggest that possible bias from this can be avoided by breaking the dependence between censoring time and X_i and by recensoring $U_i(\ )$ at the earliest possible censoring time given the treatment effect .[7] For patients that had an unfavourable treatment history (for example, control group patients that did not switch onto a beneficial new treatment; or control group patients that did switch onto a detrimental new treatment) and who experienced the event of interest close to their administrative censoring time, the event may not have been observed if they had received more favourable treatment. Therefore these patients have their survival times recensored and their events are no longer observed.[7] Recensoring is applied to all patients in groups in which switching occurred.

The RPSFTM makes a number of important assumptions – in particular, randomisation must have adequately balanced the alternative treatment arms with regard to prognostic characteristics, and the treatment effect (relative to the time the treatment is taken for) must be equal no matter when treatment is received (the "common treatment effect" assumption). Also, the accelerated failure time model must be correct, which cannot be tested.

The randomisation assumption should be reasonable in the context of an RCT, but may be called into question if sample sizes are particularly small. The "common treatment effect" assumption is more problematic. Effectively, the RPSFTM estimates an "average" treatment effect taking into account switching patients and patients initially randomised to the experimental group. If switching patients actually receive a lower treatment effect than intervention group patients, the treatment effect

estimated by the RPSFTM is likely to be an underestimate of the effect actually received in the intervention group (although this is also a function of the impact of recensoring) – and thus an underestimate of what is really required given the economic evaluation decision problem. Conversely, if switching patients happen to receive a higher treatment effect than patients in the intervention group, the RPSFTM treatment effect will be an overestimate of the effect received in the intervention group (again, depending upon the impact of recensoring). Either way, unless the "common treatment effect" assumption holds, the RPSFTM treatment effect will not deliver an unbiased estimate of the treatment effect required to address the economic evaluation decision problem. It is therefore very important to consider expert opinion on the clinical and biological plausibility of the "common treatment effect" assumption.

Recensoring represents an additional limitation of the RPSFTM method. Recensoring involves "throwing away" information, and the larger the treatment effect the more information will be discarded. Therefore in this sense the RPSFTM is inefficient, and information at the tail of the survival curve that may be important for extrapolation purposes is excluded. This is a problem particularly from the perspective of a health economist, because usually extrapolation is required to estimate mean survival times. If the treatment effect changes over time, recensoring will add to the bias associated with the RPSFTM method – for example, if the treatment effect falls over time, recensoring may lead to an over-estimate of the true (average) treatment effect.

However, a key advantage of the RPSFTM method is that it is a randomisation-based effect estimator (RBEE) as classified by White (1999).[7] By design, its significance level is equal to the significance level of an ITT analysis, and it is equal to an ITT effect estimator if there are no treatment changes or if the treatment effect is zero. The method does not rely upon the "no unmeasured confounders" assumption, described below, which affects the plausibility of observational-based approaches such as the IPCW. Whenever the ITT analysis is not significant, the RPSFTM confidence intervals will include 0. The method may therefore inspire more confidence than an observational-based approach (such as IPCW). However, this means that the confidence intervals associated with the RPSFTM treatment effect may be relatively large because the p-value from the ITT analysis is maintained. This may be regarded as a weakness of the approach, if an observational method leads to narrower confidence intervals.

Practically, there are alternative ways in which the RPSFTM method can be implemented.[1] We define these approaches as follows:

a. “On treatment” analysis. This involves a direct implementation of the RPSFTM model shown in equation (1). Under an “on treatment” approach switching is accounted for not only from the control group to the experimental treatment, but also from experimental treatment to the control group. Hence, when a patient in the experimental group discontinues treatment we assume that they have “switched” onto control group treatment. Similarly, control group patients who switch onto the experimental treatment “switch back” to the control treatment once they discontinue the experimental treatment.

   This results in a “causal” treatment effect being estimated for the experimental treatment – that is, the effect while on treatment. It is assumed that the treatment effect is lost as soon as treatment is discontinued. For an economic evaluation this causal treatment effect may not be appropriate, because the experimental group treatment discontinuation observed in the trial may reflect what would happen in real life. Hence, we require an estimate of the treatment effect associated with being randomised to the treatment group, incorporating likely treatment discontinuation rates. To obtain this from the “on treatment” RPSFTM approach, we can simply compare observed experimental group survival times to the estimated control group counterfactual survival times. Under this approach, we are estimating the treatment effect of observed treatment in the experimental group, compared to counterfactual survival times in the control group that are adjusted under the assumption that the treatment effect is only present while treatment is being taken.

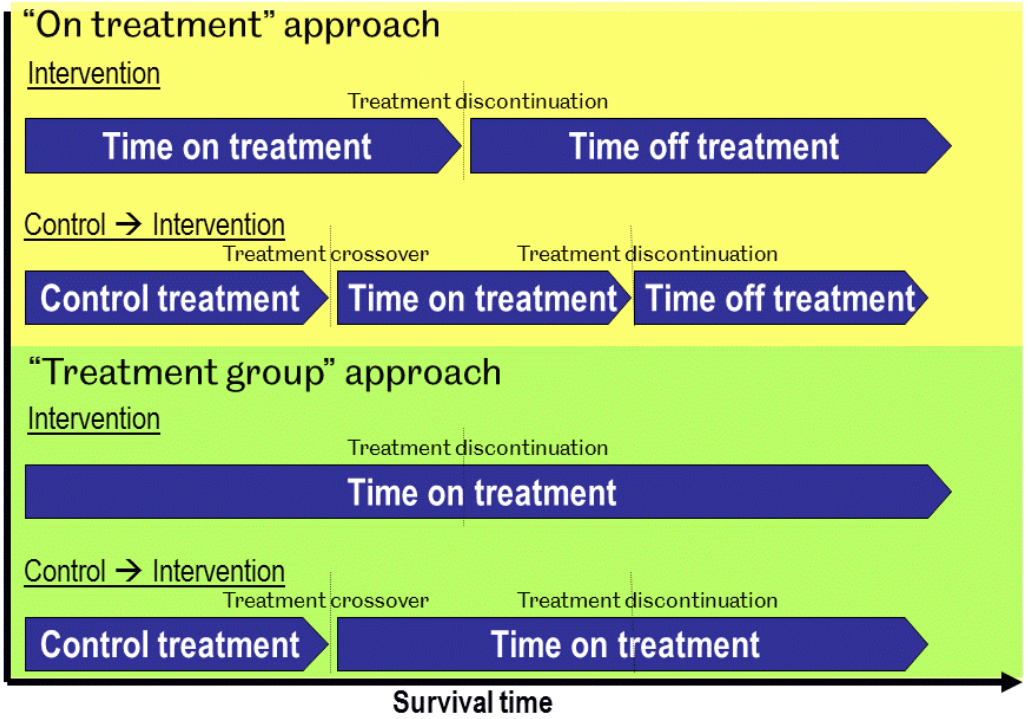
b. “Treatment group” analysis. An alternative RPSFTM application approach is the “treatment group” analysis. This may be viewed as more closely representing the type of analysis typically presented for clinical trials: the treatment effect of being randomised to the experimental group, compared to the control group (adjusted for treatment switching) is estimated, and treatment discontinuation is not specifically taken into account. Usually, in the absence of treatment switching, clinical trialists report a hazard ratio comparing the experimental group to the control group, and this is not adjusted for treatment discontinuation. Implicitly, this allows any treatment benefits associated with the new treatment that accrue after treatment discontinuation to be included in the treatment effect estimate. Similarly, if patients in the experimental group actually experience events more quickly after treatment discontinuation than patients in the control group, this will be captured within the analysis. When applying the RPSFTM method on a “treatment group” basis it is assumed that there is no switching from the experimental treatment onto the control treatment and once treatment has begun the patient remains in the “treated” group. An average treatment effect associated with being in the

experimental group is estimated, under the assumption that the treatment effect associated with being randomised to the intervention group is equal to the treatment effect received by switching patients.

However, this presents new problems with the approach, because the RPSFTM treatment effect then becomes an average treatment effect of that observed over a lifetime in the experimental group (irrespective of when treatment discontinuation occurred) and that observed from the point of switching until death in switching patients. This may "dilute" the acceleration factor in the intervention group relative to the control group if the majority of the treatment benefit is accrued while it is being taken, because the control group only begin to receive the treatment at a later time point and therefore the time between their discontinuation of treatment and death may be shorter. Hence calculating an "average" treatment effect across all patients that received the intervention may actually mean that this effect is an over-estimate of the experimental group treatment effect as the acceleration factor may actually be larger in switching patients than in intervention group patients.

Figure 2 illustrates the "on treatment" and "treatment group" approaches. The added complexity associated with the "on treatment" approach is clear. Both approaches assume that the time spent by control group patients on the control group treatment is time spent "off treatment". However, under the "treatment group" approach this is the only time spent by any patients "off treatment", whereas under the "on treatment" approach treatment discontinuation is taken into account and so intervention group patients spend a period of time after treatment discontinuation "off treatment" and control group patients spend time "off treatment" both before switching and after switching treatment discontinuation.

Figure 2: "On treatment" and "Treatment group" approaches illustrated



The treatment switching observed in the imatinib trial, combined with the data collected on these switches, means that it is difficult to satisfactorily apply the RPSFTM method for an OS analysis. As stated in Section 1 of this report, 17 control group patients switched treatments before disease recurrence. However, patients in the experimental group and the control group were also able to receive imatinib treatment after disease recurrence. Hence, for both groups there are two potential "on treatment" periods. It is important to note that this does not cause a problem for an adjusted analysis of RFS, because such an analysis only considers time up until disease recurrence, and so there can only be one period of treatment in both trial arms. However, for the OS analysis undertaking an "on treatment" approach is more complicated than is depicted in Figure 2. For instance, in the experimental group patients may receive an initial 1 year of therapy with imatinib. They then come off treatment until disease recurrence, at which point they may begin treatment with imatinib again. Under the assumption that the effect of imatinib at these two time-points is the same (relative to the time for which treatment was taken), this could be modelled by combining the two periods spent on treatment to calculate the total "on treatment" time. However, unfortunately data were not collected in the trial to allow us to determine which experimental group patients went on to receive imatinib after disease recurrence, let alone the length of this secondary treatment period. Slightly more data are available in the control group – we know which patients received

imatinib after disease recurrence in this group, but we do not know how long they took this treatment for.

This leaves three options for the application of the RPSFTM method to the imatinib dataset:

a) Adapted "on treatment" approach. We know the imatinib treatment duration given prior to disease recurrence in the experimental group, and we know that 15 of the 17 control group patients who switched before disease recurrence completed a full course of treatment – for these patients we do not have a specific treatment end date, but we can reasonably assume that they received one year of treatment. One of the remaining switchers experienced disease recurrence shortly after switching and again, although we do not have a specific treatment end date for this patient we can reasonably assume that they discontinued treatment at the time of disease recurrence. The remaining switcher discontinued treatment due to an adverse event, but we do not know the date of that event. Given no other information on this event, we may assume that they received one year of treatment.

   Given this information, we can apply the RPSFTM under the "on treatment" basis according to Figure 2 – whereby only the imatinib received prior to disease recurrence is classified as time spent "on" treatment. This requires the assumption that all other times are spent "off" treatment. This is appropriate for an adjusted analysis of RFS (because post-recurrence time periods are not included in the analysis), but is problematic for an analysis of OS given that several patients are likely to have received imatinib after disease recurrence. Only if receiving imatinib after disease recurrence did not constitute an active therapy that prolonged survival would this application of the RPSFTM be appropriate for the OS analysis. However, it may be argued that if similar proportions of patients in both treatment groups received post-recurrence imatinib this analysis may still provide a reasonable estimate of the OS treatment effect associated with pre-recurrence imatinib treatment.

b) "Treatment group (i)" approach. We are able to apply the RPSFTM on the "treatment group" basis as illustrated in Figure 2, ignoring treatment discontinuation times, and defining treatment switching as occurring when control group patients received imatinib prior to disease recurrence. Like the "on treatment" approach, this is reasonable for an analysis of RFS – with the difference between the two approaches being that the "on treatment" approach assumes that the treatment effect stops immediately upon treatment discontinuation, whereas the "treatment group" approach assumes that some benefit (or disbenefit) to treatment may remain after treatment discontinuation. Given the trial

protocol – that patients discontinue treatment after one year even if their disease has not recurred – there may be an expectation that one year of treatment is sufficient to cause a longer lasting treatment effect, and thus the "treatment group" approach may seem appropriate. However, this is essentially a clinical question, and clinical expert advice should be sought. The "treatment group (i)" approach is problematic for an OS analysis because patients in the control group who receive imatinib after disease recurrence are classified as spending all of their time "off" treatment. This would only be a valid assumption if receiving imatinib after disease recurrence did not affect survival. Both the "on treatment" and "treatment group (i)" approaches assume that the treatment effect received by pre-recurrence switchers is the same as the treatment effect received by patients initially randomised to the experimental group – an assumption that we consider in more detail in Section 4.

c) "Treatment group (ii)" approach. Under this approach we assume that patients in the control group switch and become "on" treatment whether they receive imatinib before or after disease recurrence. Again, treatment discontinuation times are ignored – once a patient receives imatinib treatment they remain "on" treatment for the remainder of their observed survival time. This approach is reasonable if we assume that the treatment effect due to imatinib is similar whether it is received pre- or post-disease recurrence – an assumption that we investigate further in Section 4. Because times of treatment discontinuation are not taken into account this approach can incorporate any post-discontinuation effects of the treatment, but this also means that the treatment pathways received after imatinib must have been valid and realistic, as these will also influence the treatment effect associated with imatinib treatment. In essence, this approach provides an estimate of the treatment effect associated with joining the imatinib treatment pathway, rather than an isolated treatment effect associated with imatinib. It should be noted that although this approach estimates the treatment effect associated with joining the imatinib treatment pathway incorporating both pre- and post-recurrence treatment switching, we only wish to adjust for pre-recurrence switching. Hence, in estimating the adjusted treatment effect associated with pre-recurrence imatinib we compare observed experimental group survival times to adjusted control group survival times, but we only adjust control group survival times for those whom switched treatment before disease recurrence. It is important to note that the "treatment group (ii)" approach is only relevant for the analysis of overall survival, because control group patients who switch onto imatinib after disease recurrence have no impact on the RFS analysis.

Finally, it is important to consider recensoring in the context of the imatinib trial. Recensoring is required when treatment switching is associated with prognosis, because this leads to counterfactual censoring times to potentially be associated with prognosis, as demonstrated by equation (3). However, if switching is not associated with prognosis, there may be no need to recensor the data. This is important to consider, because recensoring can lead to an important loss of information. In the imatinib trial, almost all control group patients who became eligible to switch on to imatinib before disease recurrence did so (17 patients switched before disease recurrence, of 18 who became eligible to switch who had not experienced disease recurrence). All patients who remained on and had been receiving placebo treatment for less than one year at the time of study unblinding (12th April 2007) were eligible to switch, and thus the opportunity to switch was influenced by the date that patients were randomised into the trial. Because almost all patients who became eligible to switch did so, it may be argued that the probability of switching was not influenced by baseline or time-dependent prognostic covariates. Because the time that patients were recruited and randomised into the trial is random (i.e. not associated with prognostic covariates), it may be argued that switching itself was random. Given this, it may not be necessary to recensor analyses that adjust specifically for pre-recurrence switching in the imatinib trial. Therefore, because the RPSFTM options (a) and (b) described above only define switching as occurring when patients switch before disease recurrence, recensoring may not be required. However, for RPSFTM option (c) recensoring will be required, because post-recurrence switching is also included in the analysis, and this switching may have been influenced by prognostic factors.

2.2 Iterative Parameter Estimation Algorithm

Branson and Whitehead (2002) extended the RPSFTM method using parametric methods, developing a novel iterative parameter estimation (IPE) procedure.[8] The same type of accelerated failure time model is used, but a parametric failure time model (for example a Weibull, log-logistic, log normal) is fitted to the original, unadjusted ITT data to obtain an initial estimate of . The observed failure times of switching patients are then re-estimated using and the counterfactual survival time model presented in equation (1), and the treatment groups are then compared again using a parametric model. This will give an updated estimate of , and the process of re-estimating the observed survival times of switching patients is repeated. This iterative process is continued until the new estimate for is very close to the previous estimate (the authors suggest within 10^{-5} of the previous estimate but offer no particular rationale for this), at which point the process is said to have converged.[8]

Branson and Whitehead (2002) initially suggested a more limited version of recensoring than that suggested by White *et al.* (1999). However, this was subsequently shown to be prone to bias and thus full recensoring is required, if switching is associated with prognosis.[9]

Branson and Whitehead (2002) state standard errors and confidence intervals for the treatment effect could be taken from the final regression in the IPE process, or bootstrapping could be used.[8] However, the authors state that the standard errors from the final regression will be underestimates because the covariance matrix from the final iteration of the IPE algorithm does not take into account the fact that control arm patients have had their survival times transformed by the algorithm – therefore bootstrapping is preferable (although in the context of re-applying the IPE method to bootstrapped trial data this can be computationally expensive). The authors also recommend that the IPE estimate and confidence interval is accompanied by the ITT p-value because as for the RPSFTM method the significance level remains the same as that associated with the ITT analysis.[8]

The IPE procedure makes similar assumptions to the RPSFTM method – for example the randomisation assumption is made, as is the "common treatment effect" assumption. An additional assumption made by Branson and Whitehead (2002) is that survival times follow a parametric distribution, and thus they state that it is important to identify suitable parametric models using tools such as log-cumulative hazard plots. Hence similar advantages and disadvantages exist for the IPE method as compared to the RPSFTM method, with the addition of possible disadvantages if a suitable parametric distribution cannot be identified.

Branson and Whitehead (2002), Morden *et al.* (2011) and Latimer *et al.* (2014) conducted simulation studies to assess the performance of the IPE method, and found that the method produces very low levels of bias when the assumptions made by the method (in particular, the "common treatment effect" assumption) hold.[1,2,8] Morden *et al.* (2011) and Latimer *et al.* (2014) found that the method's performance was very similar to the RPSFTM, which is to be expected as the methods are similar other than in their estimation procedure.

With regard to the application of the IPE method to the imatinib dataset, it can be applied in exactly the same way as the RPSFTM – i.e. using the "treatment group" and "on treatment" approaches described in Section 2.2. Therefore, the application of the IPE method suffers from the same limitations as the RPSFTM method in this case study.

2.3 Inverse Probability of Censoring Weights

An inverse probability of censoring weighted Cox model represents a type of Marginal Structural Model (MSM) that is relevant for the context of treatment switching. The method can be used as an approach for addressing any kind of informative censoring. In the context of treatment switching the standard application of involves censoring patients at the point at which they switch.[10] Remaining patients have their observations weighted taking into account their baseline and time-dependent covariate values compared to those of switching patients. The Cox partial likelihood estimator of the hazard ratio is then replaced by its IPCW version.

The IPCW method involves weighting the contribution to the survivor function of a subject at risk at time t by the inverse of an estimate of the conditional probability of having remained uncensored until time t. Robins and Finkelstein (2000) recommend the use of "stabilised" weights, as these are shown to be more efficient.[10] Unstabilised weights are simply the inverse of the conditional probability of having remained uncensored until time t conditional on baseline and time-dependent covariates, whereas stabilised weights are the conditional probability of having remained uncensored until time t given baseline covariates, divided by the conditional probability of having remained uncensored until time t given baseline and time-dependent covariates. The stabilised weight will be equal to 1 for all t if the histories of the included prognostic factors for failure do not impact upon the hazard of censoring at t – thus there would be no informative censoring and treatment switching would be random.[10]

Formally, the stabilised weights applied to each individual for time interval (t), as specified by Hernan *et al.* (2001) are:[11]

$$\prod \frac{[\qquad\qquad\qquad\qquad]}{[\qquad\qquad\qquad\qquad]} \quad (4)$$

Where is an indicator function demonstrating whether or not informative censoring (switching) had occurred at the end of interval k, and denotes censoring history up to the end of the previous interval (k-1). denotes an individual's treatment history up until the end of the previous interval (k-1), and V is an array of an individual's baseline covariates. denotes the history of an individual's time-dependent covariates measured at or prior to the beginning of interval k. Hence the numerator of (4) represents the probability of an individual remaining uncensored (not switched) at the end of interval k given that that individual was uncensored at the end of the previous interval (k-1), conditional on baseline characteristics and past treatment history. The denominator represents that same probability conditional on baseline characteristics, time-dependent characteristics and past treatment history. When the cause of informative censoring is

treatment switching, past treatment history is removed from the model because as soon as switching occurs the individual is censored.

The probabilities required to estimate the stabilised weight can be estimated by fitting logistic regression models with informative censoring (switching) as the dependent variable, and with baseline covariates included as independent variables for calculating the numerator of the weight, and baseline and time-dependent covariates included as independent variables for the denominator.[11]

The IPCW adjusted Cox hazard ratio (HR) can be estimated by fitting a time-dependent Cox model to the dataset in which switching patients are artificially censored. The model includes baseline covariates and uses the time-varying stabilised weights for each patient and each time interval.

The IPCW approach makes use of a large amount of covariate data in order to inform estimates of survival adjusting for informative censoring (switching) and owing to this the method is potentially more powerful than a standard ITT or randomisation-based (RPSFTM or IPE) analysis. However, this is dependent on covariate information being available on baseline and time-varying covariates that are prognostic of switching, and may not be the case if there is much uncertainty in the data and/or the trial dataset is relatively small.

Related to this, the IPCW method is reliant upon the untestable "no unmeasured confounders" assumption, which means that data must be available on all baseline and time-dependent covariates that influence the probability of switching or death. Hence, the IPCW method may be described as "observational-based", although once weightings have been applied groups are compared as randomised. In addition, models for switching and death must be correctly specified,[10] and the approach cannot work if there are any covariates which ensure (that is, the probability equals 1) treatment switching will or will not occur.[11]

A further potential disadvantage of the IPCW approach compared to the RPSFTM approach is that it ignores potentially useful information from patients who switch and thus are censored from the analysis. However, from the economic evaluation perspective this may actually be perceived as an advantage, as it means that the IPCW method specifically addresses the decision problem – that is, it specifically estimates the treatment effect in patients randomised to the experimental treatment.

The IPCW method has a strong theoretical basis and has been shown to be unbiased when its assumptions hold.[10] However applications of the method have highlighted problems associated with data availability,[12,13] and simulation studies have shown that the method is prone to bias for several reasons.[1,14] Latimer *et al.* (2014) note that the IPCW method is particularly prone to high

levels of bias when very high proportions of patients switch.[1] Howe *et al.* (2011) demonstrate that the IPCW method may result in biased estimates of survival when sample sizes are very small (n=50), selection bias is very strong (with a sample size of 500), and if there exist unmeasured confounders.[14] The problems associated with the IPCW approach will be exacerbated in circumstances where few events have been observed, as modelling will be more uncertain.

In the context of the imatinib trial, a standard application of the IPCW is problematic, because almost all control group patients who became eligible to switch onto imatinib prior to disease recurrence did so. It is in situations such as these – where an extremely high proportion of patients who become at-risk of switching do switch – that the IPCW has been shown to result in high levels of bias,[1] largely because it is not possible to accurately model the probability of switching. However, in the case of the imatinib trial a simpler application of the IPCW method may be applicable, and may be prone to minimal bias. As described in Section 2.1, the pre-disease recurrence switching observed in the imatinib trial might be considered to be random, since all except one patient who became eligible to switch treatments before disease recurrence did so, and because switching eligibility was essentially a function of randomisation date. Because of this, it may be argued that a simple per-protocol censoring approach – whereby pre-recurrence switchers are censored at the time of switch – may be suitable, because this should not result in selection bias. However, an IPCW analysis may still be preferable, to adjust for any chance differences in baseline characteristics between control group patients who switched pre-recurrence and those that did not, and to take account of the time at which switchers switched. In fact, such an analysis would represent a useful check on the simple per-protocol censoring approach – if the two provide similar results, and if IPCW weights are close to 1.0, we can be more confident in the results. An IPCW applied in this way uses the entire control group when estimating stabilised weights, rather than only those patients who actually became eligible to switch. Hence, because only 17 patients out of 98 in the control group switched before disease recurrence the switching proportion is quite low (17.3%), and therefore the imatinib trial does not represent a situation in which we would expect high levels of bias associated with the IPCW. Added to this, we do not require time-dependent covariates to be included in our model (only time since randomisation), and thus the scope for bias due to unmeasured confounders and incorrectly specified models is substantially reduced. Finally, because the IPCW censors switchers it avoids complications around classifications of “on treatment” and “off treatment”: The IPCW simply provides an estimate of what the treatment effect associated with pre-recurrence treatment with imatinib would have been in the absence of any pre-recurrence control group switching in the imatinib trial.

2.4 Two-stage methods

Recent research has suggested that two-stage methods may be useful when addressing the treatment switching problem.[1] These involve estimating the treatment effect specific to switching patients and then using this to "shrink" the survival times of switchers to arrive at a counterfactual dataset. While this method has been shown to perform well in simulation studies it requires that there is a "secondary baseline" within the trial, after which treatment switching becomes possible.[1] This secondary baseline must be disease oriented, such that the prognosis of patients can be classified as similar at this time-point. An example of such a time-point in the context of the imatinib trial is disease recurrence. If treatment switching was only allowed after disease recurrence, and if we wished to adjust for switching that occurred beyond recurrence, we could use the time of disease recurrence as a secondary baseline for each patient, and then we could have estimated the effect of treatment after this time-point. However, in the imatinib trial we wish to adjust for switching that occurred before disease recurrence, not after, and thus a suitable secondary baseline does not exist – patients became eligible to switch based upon a calendar time, not a time-point related to disease progression. Hence, we cannot apply two-stage methods to the imatinib case study presented in this report.

2.5 Methods Summary

In the presence of treatment switching the RPSFTM, IPE algorithm and IPCW methods represent potentially appropriate approaches for adjusting estimates of the treatment effect. However, each method has important limitations and these should be taken into account when considering which approach is likely to be optimal for specific trial datasets.

For the RPSFTM and IPE methods it is usually most important to consider the plausibility of the randomisation and "common treatment effect" assumptions. For the IPCW method the plausibility of the "no unmeasured confounders" assumption must be considered, as well as the likelihood that the treatment switching and survival processes can be accurately modelled. Previous research has shown that RPSFTM and IPE algorithm methods produce very little bias provided the "common treatment effect" assumption holds, but otherwise these methods are prone to substantial bias, depending upon the extent to which the treatment effect differs between switching patients and patients initially randomised to the experimental group.[1,2] It has also been shown that the IPCW method is prone to moderate levels of bias in relatively small sample sizes (n=500) even if the "no unmeasured confounders" assumption holds.[2,14]

However, the specifics of the trial in question and the data that are available must also be considered. As described in Section 2.1, the RPSFTM and IPE counterfactual survival model does not represent a good fit to the imatinib trial, owing to the potential for two separate exposures to imatinib during the trial (one pre-recurrence, and one post-recurrence) combined with a lack of data on treatment durations in the post-recurrence period. In Section 2.1 we described three approaches that may be taken to apply the RPSFTM and IPE methods in the context of the imatinib trial, but none are completely satisfactory due to the assumptions that they require. On the other hand, the IPCW method seems to provide a good fit to the treatment switching problem encountered in the imatinib trial, for a number of reasons: switching appears likely to have been random with regard to prognostic covariates meaning we do not need to model using time-dependent covariates (apart from time since randomisation); the switching proportion is relatively low as a proportion of the control group as a whole; and, we can avoid making assumptions about the relative effectiveness of imatinib pre- and post-disease recurrence.

2.6 Methods Applied in this Report

In line with the above discussion we have applied the following "complex" treatment switching adjustment methods to RFS and OS data:

a. RPSFTM
    - "On treatment" analysis using log-rank test (option (a) in Section 2.1)
    - "Treatment group (i)" analysis using log-rank test (option (b) in Section 2.1)
    - "Treatment group (ii)" analysis using log-rank test (option (c) in Section 2.1, only applied to OS analysis)
b. IPE algorithm
    - "On treatment" analysis (option (a) in Section 2.1) using Weibull model
    - "Treatment group (i)" analysis (option (b) in Section 2.1) using Weibull model
    - "Treatment group (ii)" analysis (option (c) in Section 2.1, only applied to OS analysis) using Weibull model
c. IPCW. Applied as described in Section 2.3

Note, for the RPSFTM and IPE "on treatment" and "treatment group (i)" analyses we have applied the methods both with and without recensoring.

In addition, simple per-protocol methods have been applied in order to provide a range of OS treatment effect estimates for completeness, comparison and discussion. These include:

a. Censor switching patients

b. Exclude switching patients

These are of particular interest in this case study, since we believe that the treatment switching observed in the trial may have been random. Hence, we may expect these analyses to provide similar results to the potentially more appropriate complex adjustment methods. In this case study, given the nature of the trial and data available, *a priori* we believe the IPCW may represent the most appropriate analyses for RFS and OS. However, the RPSFTM/IPE "on treatment" and "treatment group (i)" analyses may represent additional plausible analyses for RFS, and the RPSFTM/IPE "treatment group (ii)" may represent additional plausible analyses for OS.

In our application of the IPCW method we included baseline covariates for region, age, sex, time since resection, tumour size, whether the tumour was located in the stomach, performance status, resection margin, and mitoses. These represent the prognostic covariates that were measured at baseline in the imatinib trial and which were used within exploratory adjusted analyses in the clinical study report. The only potentially prognostic covariate that was excluded identified the mutation category of the patient, which was missing for several patients, and which was EXON 11 for the vast majority of patients (71.6%) for whom data were available. Of 14 switching patients for whom data on mutation category was available, 9 (64.3%) had an EXON 11 mutation. We used Stata Version 11.2 [15] to carry out our analysis, and used Stata's rcsgen command to generate restricted cubic splines with five knots for our IPCW application. The code used to apply the IPCW method was similar to that presented by Fewell *et al.*[16]

We used the strbee Stata program to apply the RPSFTM and IPE algorithm methods.[17] In our application of the RPSFTM we used the log-rank test and we tested the method both with and without recensoring. In our application of the IPE algorithm, the Weibull distribution was used for the iterative estimation of the treatment effect. Other parametric accelerated failure time models could be used and thus model uncertainty is introduced. We have essentially run the IPE algorithm analyses as a check on the RPSFTM analyses – we expect these to give very similar results. Given this, and given the relative results of the RPSFTM and IPE analyses described below, we decided not to take the additional step of running bootstrapped iterations to estimate IPE algorithm confidence intervals.

3. Results

In this section the results of our switching adjustment analyses are presented. Further details on these results are presented in Appendices 1, 2 and 3. Appendix 1 presents plots of the log-rank Z statistic for the RPSFTM analyses, and Appendix 2 presents Kaplan-Meier plots for the RPSFTM and IPE analyses. Appendix 3 presents weighted Kaplan-Meier curves calculated using the IPCW method.

3.1 Results: RFS

In the imatinib trial 201 patients were in the high-risk patient group focussed upon in this report. 103 of these were randomised to the imatinib group, and 98 were randomised to the placebo group. Of the 98 randomised to the placebo group, 17 (17.3%) switched onto imatinib before they had experienced disease recurrence and after unblinding of the trial on 12th April 2007. At the time of the data-cut for the 5-year analysis, 7 (41.2%) of the 17 switchers had experienced disease recurrence, compared to 53 (65.4%) of control group non-switchers. 50 (48.5%) of the 103 imatinib group patients had experienced disease recurrence by the time of the 5-year data cut.

Table 1 reports the results of switching adjustment analyses undertaken on the imatinib RFS dataset. The hazard ratios (HR) presented are of primary interest, because they represent the adjusted estimate of the treatment effect for each adjustment method. Acceleration factors (AFs) are presented for the RPSFTM and IPE methods because these methods use accelerated failure time models and so provide estimates of the treatment effect in terms of an AF, from which the adjusted HR is derived. For the RPSFTM and IPE "on treatment" analyses, the AF represents the causal treatment effect received while "on" treatment – that is, the treatment effect associated with full treatment (with no discontinuation), under the assumption that a treatment benefit is only received while a patients is actually taking imatinib. As explained in Section 2.1, we are interested in the treatment effect associated with imatinib, taking into account that treatment discontinuation will occur. Therefore the HR presented for the "on treatment" analyses compare *observed* experimental group survival to the adjusted control group survival times.

Table 1: RFS analysis results

Analysis	Method type	Description	HR			AF		
			Point estimate	lower 95% CI	Upper 95% CI	Point estimate	lower 95% CI	Upper 95% CI
1	'Baseline', unadjusted	Intention to treat analysis (to provide a 'baseline' analysis	0.61	0.42	0.89	-	-	-
2	Naive	Per protocol: Censor switchers at time of switch	0.52	0.35	0.77	-	-	-
3	Naive	Per protocol: Exclude switchers	0.50	0.34	0.73	-	-	-
4	Complex	RPSFTM 'on treatment' analysis, with recensoring	0.10	0.01	0.82	196.97	94.00	-
5	Complex	RPSFTM 'on treatment' analysis, without recensoring	No convergence					
6	Complex	RPSFTM 'treatment group (i)' analysis, with recensoring	0.25	0.09	0.71	2.93	1.46	5.04
7	Complex	RPSFTM 'treatment group (i)' analysis, without recensoring	0.52	0.32	0.85	2.44	1.24	4.30
8	Complex	IPE 'on treatment' analysis, with recensoring	0.00*	-	-	204.64	-	-
9	Complex	IPE 'on treatment' analysis, without recensoring	0.57*	-	-	1.8E+60	-	-
10	Complex	IPE 'treatment group (i)' analysis, with recensoring	0.26	-	-	2.80	-	-
11	Complex	IPE 'treatment group (i)' analysis, without recensoring	0.55	-	-	1.80	-	-
12	Complex	IPCW	0.50	0.31	0.78	-	-	-

*Did not converge.

Note: results given to 2 decimal places. Note, CIs are not presented for the IPE analyses as bootstrapping has not been completed.

The first thing to note from these results is that the RPSFTM and IPE "on treatment" analyses did not perform as desired. The RPSFTM "on treatment" analysis that did not incorporate recensoring did not converge at all, and the version that incorporated recensoring failed to estimate an upper confidence interval for the AF. This is because no AFs above the value of 1 can be confidently excluded which is likely to be due to the low event numbers observed while the experimental treatment was being taken (treatment was usually discontinued before recurrence occurred). The IPE "on treatment" analysis excluding recensoring results in a more plausible estimate of the HR, but neither this analysis or the IPE "on treatment" analysis including recensoring converged, and the seemingly plausible result in Analysis 9 may be due to chance, as the acceleration factor (AF) estimated by the method was incredibly high. These results should not be relied upon, and we consider the results of "on treatment" analyses no further.

The RPSFTM and IPE "treatment group" analyses give very different results depending upon whether or not recensoring is incorporated. The Kaplan-Meier plots for these analyses presented in Appendix 2 demonstrate that recensoring has a very important effect in these analyses, with the survival times of patients who did not die during the trial being recensored at much earlier time-points, causing the adjusted Kaplan-Meier curve to drop towards zero very quickly because the longer-term control

group events become death events rather than censored events. This results in a substantial loss of information on patients who lived until the end of the trial follow-up, and the resulting treatment effect estimates are implausible. Given this, and given our arguments in Section 2 that suggest that pre-disease recurrence switching in the imatinib trial was likely to be random, we believe that for the RFS analysis RPSFTM and IPE "treatment group (i)" analyses that are not subject to recensoring are likely to be more appropriate.

When using the RPSFTM or IPE adjustment methods the estimation procedure can be checked by comparing the hazard ratio estimated between control group and experimental group counterfactual survival times – that is, the HR between the groups if no patients in either group had received treatment. If the estimation procedure has worked well, this value should be very close to 1.0 (this reflects that the method is reliant upon randomisation having created well balanced treatment groups). For the RPSFTM "treatment group (i)" analysis without recensoring (Analysis 7) this was 0.9998, and for the corresponding IPE analysis (Analysis 11) it was 0.8321, suggesting that the RPSFTM method outperformed the IPE method.

In Appendix 1 graphs plotting the log-rank Z statistic compared to values of psi ( ) are shown for the RPSFTM analyses. equals –ln(acceleration factor). The RPSFTM estimate of is taken as the value for which the log-rank Z statistic is equal to 0. The upper and lower confidence intervals are taken as the value of for which the log-rank Z statistic is equal to ±1.96. Plotting the log-rank Z statistic for each value of helps demonstrate whether there is a unique solution to Z( )=0.[17] The graph for the RPSFTM "treatment group" analysis without recensoring (Analysis 7) indicates that this is the case, with the plot forming close to a straight downward line.

It is noteworthy that the confidence intervals associated with the RPSFTM "treatment group" analysis without recensoring are wide, as they incorporate the p-value associated with the ITT analysis. We have not presented the confidence intervals for the IPE analyses as this requires bootstrapping, which seems unnecessary given that the RPSFTM analyses appear more robust.

The IPCW analysis provides an adjusted HR of 0.50 (95% Confidence Interval 0.31 – 0.78), which is similar to the simple per-protocol censoring analysis (HR 0.52 (95% CI 0.35 – 0.77)). The mean stabilised weight applied within the analysis was 1.00, with the minimum weight being 0.15 and the maximum being 1.47. The low range of these weights demonstrates that no individual patients are being allocated particularly high weights, which means that there is reduced scope for error in the analysis. This, combined with the fact that the IPCW results are very similar to the per-protocol censoring approach supports the hypothesis that the switching observed may have been random.

The IPCW analysis may still be preferred over the simple per-protocol censoring analysis because it adjusts for any minor differences in baseline characteristics between control group switchers and control group non-switchers, and accounts for the time since randomisation at which switching occurred.

The potential validity and robustness of the IPCW and simple per-protocol censoring analyses are further strengthened by the fact that the RPSFTM analysis that appears to have performed most appropriately (RPSFTM "treatment group (i)" without recensoring (Analysis 7)) resulted in a similar adjusted HR of 0.52 (95% CI 0.32 – 0.85). Similarly, in the presence of random switching, the per-protocol analysis that excludes switchers (Analysis 3) would also provide an unbiased estimate of the adjusted treatment effect. This analysis provided an HR of 0.50 (95% CI 0.34 – 0.73), again, similar to the estimates provided by the simple per-protocol censoring analysis, the IPCW analysis, and the RPSFTM "treatment group (i)" without recensoring analysis. We believe that the IPCW adjusted HR of 0.50 (95% CI 0.31 – 0.78) provides a robust estimate of the treatment effect of imatinib on RFS, adjusted to take account of pre-disease recurrence treatment switching. However, to support this we conduct more analysis on the plausibility of the key assumptions associated with the IPCW and RPSFTM analyses in Section 4 of this report.

3.2 Results: OS

At the time of the data-cut for the 5-year analysis, 17 (21.0%) of the 81 control group non-switchers had died, compared to 0 (0.0%) of the 17 control group switchers. 17 (16.5%) of the 103 imatinib group patients had died. 36 (60.0%) of the 60 control group patients who had experienced disease recurrence had received imatinib post-disease recurrence, but data on post-recurrence imatinib treatment was missing in 18 control group patients who had experienced disease recurrence. Data are not available on when post-disease recurrence treatment with imatinib started, and thus for the "treatment group (ii)" analysis we assume this treatment began on the date that each patient was unblinded. For the 19 control group patients for whom data on post-recurrence imatinib treatment is missing, we assumed that treatment switching did not occur.

Table 2 reports the results of switching adjustment analyses undertaken on the imatinib OS dataset.

Table 2: OS analysis results

Analysis	Method type	Description	HR			AF		
			Point estimate	lower 95% CI	Upper 95% CI	Point estimate	lower 95% CI	Upper 95% CI
13	'Baseline', unadjusted	Intention to treat analysis (to provide a 'baseline' analysis	0.93	0.47	1.83	-	-	-
14	Naive	Per protocol: Censor switchers at time of switch	0.79	0.40	1.55	-	-	-
15	Naive	Per protocol: Exclude switchers	0.79	0.40	1.54	-	-	-
16	Complex	RPSFTM 'on treatment' analysis, with recensoring	0.10	0.05	0.20	8.36	0.22	-
17	Complex	RPSFTM 'on treatment' analysis, without recensoring	0.91	0.47	1.80	1.53	0.22	-
18	Complex	RPSFTM 'treatment group (i)' analysis, with recensoring	0.52	0.00	225.39	1.31	0.66	2.71
19	Complex	RPSFTM 'treatment group (i)' analysis, without recensoring	0.92	0.42	2.00	1.31	0.57	1.81
20	Complex	RPSFTM 'treatment group (ii)' analysis, without recensoring	0.82	0.42	1.60	1.90	0.73	3.92
21	Complex	IPE 'on treatment' analysis, with recensoring	0.09	-	-	4.54	-	-
22	Complex	IPE 'on treatment' analysis, without recensoring	0.92	-	-	1.47	-	-
23	Complex	IPE 'treatment group (i)' analysis, with recensoring	0.42	-	-	1.44	-	-
24	Complex	IPE 'treatment group (i)' analysis, without recensoring	0.92	-	-	1.05	-	-
25	Complex	IPE 'treatment group (ii)' analysis, without recensoring	0.81	-	-	2.02	-	-
26	Complex	IPCW	0.76	0.36	1.62	-	-	-

Note: results given to 2 decimal places. Note, CIs are not presented for the IPE analyses as bootstrapping has not been completed.

As for the RFS analysis, the first thing to note from these results is that the RPSFTM and IPE "on treatment" analyses did not perform as desired. The RPSFTM "on treatment" analyses (with and without recensoring) partly converged, but did not produce upper confidence intervals for the AF. This is because no AFs above the value of 1 can be confidently excluded which is likely to be due to the low event numbers observed while the experimental treatment was being taken (treatment was usually discontinued before recurrence occurred), combined with the fact that there is little evidence against AF values of greater than 1 due probably to the low event numbers observed. This demonstrates the problems associated with analysing small datasets with few observed events, and in this instance, this problem is exacerbated for "on treatment" analyses. Although both the RPSFTM and the IPE "on treatment" analyses that excluded recensoring resulted in plausible estimates of the adjusted HR, we believe that these should not be relied upon given the estimation problems associated with these analyses.

As was the case for the RFS analysis, the RPSFTM and IPE "treatment group (i)" analyses gave very different results depending upon whether or not recensoring was incorporated. The Kaplan-Meier

plots for these analyses presented in Appendix 2 demonstrate that recensoring again has an important impact, again representing an important loss of information in patients who remained alive at the end of the trial follow-up. Given this, and given our arguments in Section 2 that suggest that pre-disease recurrence switching in the imatinib trial was likely to be random, we believe that for the OS analysis RPSFTM and IPE "treatment group (i)" analyses that are not subject to recensoring are likely to be more appropriate. For these, the corresponding RPSFTM and IPE analyses provide very similar estimates of the adjusted HR (0.92). The hazard ratio estimated between control group and experimental group counterfactual survival times was 1.001 for the RPSFTM and 0.991 for the IPE, suggesting that the RPSFTM estimation procedure performed marginally better. The graph for the RPSFTM "treatment group (i)" analysis without recensoring (Analysis 19) illustrates a plot forming close to a straight downward line, again suggesting that the estimation procedure was successful.

However, in Section 2 we discussed the limitations of the "treatment group (i)" approach applied to the imatinib trial OS data, because in this analysis patients in the control group who receive imatinib after disease recurrence are classed as having never received treatment. Clearly, this could confound our results, and so we also ran the "treatment group (ii)" analysis, in which patients in the control group who received imatinib after disease recurrence were defined as switchers (and hence it was assumed that the relative treatment effect received by these patients was the same as the treatment effect received by control group patients who switched before disease recurrence, and the same as patients initially randomised to the experimental group). Although this "common treatment effect" assumption may be problematic, we believe that for OS, this version of the RPSFTM and IPE analyses fits the imatinib trial characteristics better than the "treatment group (i)" and "on treatment" analyses. The RPSFTM and IPE "treatment group (ii)" analyses provided estimates of the adjusted HR of 0.82 (95% CI 0.42 – 1.60) and 0.81 respectively. For these analyses we incorporated recensoring, because the switching onto imatinib that occurred after disease recurrence is unlikely to have been random – unlike the switching that occurred before recurrence. The Kaplan-Meier plots shown in Appendix 2 show that recensoring has less of an impact on these analyses, with no sudden drops in the adjusted Kaplan-Meier curve towards the end of follow-up. It is noteworthy that the confidence intervals associated with the RPSFTM "treatment group (ii)" analysis are wide, as they incorporate the p-value associated with the ITT analysis. Again, we have not presented the confidence intervals for the IPE analyses as this requires bootstrapping, which seems unnecessary given that the RPSFTM analyses appear more robust.

The IPCW analysis provides an adjusted HR of 0.76 (95% Confidence Interval 0.36 – 1.62), which is similar to the simple per-protocol censoring analysis (HR 0.79 (95% CI 0.40 – 1.55)). The mean stabilised weight applied within the analysis was 1.00, with the minimum weight being 0.05 and the maximum being 1.63. As for the RFS analysis, the low range of these weights demonstrates that no individual patients are being allocated particularly high weights, which means that there is reduced scope for error in the analysis. This, combined with the fact that the IPCW results are very similar to the per-protocol censoring approach supports the hypothesis that the switching observed may have been random. The IPCW analysis may still be preferred over the simple per-protocol censoring analysis because it adjusts for any minor differences in baseline characteristics between control group switchers and control group non-switchers, and accounts for the time since randomisation at which switching occurred.

As for the RFS analysis, the potential validity and robustness of the IPCW and simple per-protocol censoring analyses are further strengthened by the fact that the RPSFTM analysis that appears to best fit the imatinib trial characteristics (RPSFTM "treatment group (ii)" (Analysis 20)) resulted in an reasonably similar adjusted HR of 0.82 (95% CI 0.42 – 1.60). Similarly, in the presence of random switching, the per-protocol analysis that excludes switchers (Analysis 3) would also provide an unbiased estimate of the adjusted treatment effect. This analysis provided an HR of 0.79 (95% CI 0.40 – 1.54), again, similar to the estimates provided by the simple per-protocol censoring analysis, the IPCW analysis, and the RPSFTM "treatment group (ii)" analysis. We believe that the IPCW adjusted HR of 0.76 (95% CI 0.36 – 1.62) provides a robust estimate of the treatment effect of imatinib on OS, adjusted to take account of pre-disease recurrence treatment switching. However, to support this we conduct more analysis on the plausibility of the key assumptions associated with the IPCW and RPSFTM analyses in Section 4 of this report.

4. Key issues

In Section 2 we considered the validity of applying RPSFTM, IPE and IPCW methods to the imatinib trial dataset. In this section we revisit these issues, and present additional analyses to provide further information regarding the appropriateness of the alternative adjustment methods.

4.1 Modelling approaches

As discussed in Section 2, applying the RPSFTM and IPE methods to the imatinib dataset in an appropriate manner is difficult due to the characteristics of the trial and the data available. This is particularly the case for an analysis of OS. The counterfactual survival model utilised by both methods assumes that at any time-point during the trial each patient is either "on" or "off" treatment. However, the trial allowed imatinib to be given after disease recurrence, and thus experimental group patients, and control group switchers, had two potential "on" treatment periods. In itself this is not a problem, because these two periods could be combined to calculate one total "on" treatment period. However, data were not collected on post-recurrence imatinib treatment in the experimental group, and while these data were collected in the control group, the duration of this treatment was not collected. Hence a standard RPSFTM or IPE "on treatment" analysis cannot be undertaken. The situation would not be improved by assuming that post-recurrence imatinib constituted a different, independent treatment, because we would still need to know who took this treatment and for how long, and because previous research has shown that multiparmater versions of RPSFTM models – estimating the effects of more than one treatment – lack power and do not provide useful results.[7,13,18]

We were restricted to running three variations of the RPSFTM / IPE methods. First, an "on treatment" approach was applied, where only pre-disease recurrence treatment with imatinib was defined as time spent "on" treatment. This is problematic for an analysis of OS, because it classifies time spent on imatinib after disease recurrence as time "off" treatment – an assumption that would only be valid if post-recurrence treatment with imatinib did not prolong survival. However, even if this assumption did not hold this analysis may be argued to lead to a reasonable estimate of the adjusted OS treatment effect associated with pre-recurrence imatinib if similar proportions in the control and intervention groups received post-recurrence imatinib treatment. Unfortunately data were not collected that would allow us to test this. We ran this "on treatment" analysis both with and without incorporating recensoring, because we envisaged that recensoring may have a substantial impact on the results of the analysis, and also because we hypothesised that recensoring may not be necessary because treatment switching appears likely to have been random, and

unrelated to prognostic covariates. In practice, neither the RPSFTM or IPE "on treatment" analyses with or without recensoring performed well in either our OS or our RFS analyses, with imperfect convergence. Hence these analyses cannot be relied upon.

The second variation of the RPSFTM and IPE methods that we applied were "treatment group (i)" analyses. Under this analysis treatment switching was defined as occurring when patients in the control group switched onto imatinib before disease recurrence. Patients randomised to the imatinib group are classified as being "on" treatment for the duration of the trial, and patients in the control group who switch onto imatinib pre-disease recurrence are classified as remaining "on" treatment until the end of follow-up. This ignores treatment discontinuation times, but is similar in concept to standard Cox regression models commonly used to estimate the treatment effect associated with being randomised to the experimental group in a clinical trial. However, while this approach may be reasonable for an analysis of RFS, it is more problematic for an OS analysis, because for control group patients who do not switch onto imatinib prior to disease recurrence, it regards post-recurrence treatment with imatinib as time "off" treatment. This would only be a valid assumption if post-recurrence treatment with imatinib did not prolong survival. If post-recurrence treatment with imatinib does prolong survival, this assumption is more problematic than the corresponding assumption for the "on treatment" analysis, because in the "treatment group" analysis time spent on post-recurrence imatinib is implicitly classified as time "on" treatment in the experimental group and in switchers, but is classified as time "off" treatment in control group non-switchers. Hence, we believe that the "treatment group (i)" analyses may be reasonable for our analysis of RFS, but is less suitable for an analysis of OS. Our results presented in Section 3.1 show that for RFS the "treatment group (i)" RPSFTM and IPE analyses produced what appear to be plausible adjusted HR estimates, provided recensoring was not undertaken (HR 0.52, 95% CI 0.32 – 0.85 for the RPSFTM). Given that switching appears likely to have been random, we believe these results are reasonable, although this is dependent upon the validity of the "common treatment effect" assumption – that is, the treatment effect received by control group patients who switched onto imatinib prior to disease recurrence was the same as (or at least similar to) the treatment effect received by patients initially randomised to the experimental group. We will investigate this assumption further in Section 4.2, below. We believe that these results are plausible because they provide similar estimates to those provided by the IPCW analysis (HR 0.50, 95% CI 0.32 – 0.78), and the simple per-protocol censoring analysis (HR 0.52, 95% CI 0.35 – 0.77), which we will discuss in more detail below. For our analysis of OS, the "treatment group (i)" analyses provided results that were further away from those provided by analyses that we believe are more likely to be

appropriate (IPCW and simple per-protocol censoring). This is likely to be due to the post-recurrence imatinib treatment classification problems associated with this analysis.

The third variation of the RPSFTM and IPE methods that we applied were "treatment group (ii)" analyses, which were only applicable to our analysis of OS. This analysis addressed the problem associated with the classification of post-recurrence treatment with imatinib in the control group. Treatment switching was defined as occurring when control group patients switched onto imatinib, whether this occurred before or after disease recurrence. Given the active nature of imatinib treatment, we believe that this analysis is more appropriate than the "treatment group (i)" analysis for OS, although this is dependent upon the treatment effect received by patients who switch onto imatinib after disease recurrence being similar to the treatment effect received by patients who received imatinib before disease recurrence – we will investigate this assumption further in Section 4.3, below. Our results presented in Section 3.2 demonstrate that this method produced adjusted OS HR estimates (0.82, 95% CI 0.41 – 1.60 for the RPSFTM) that were quite similar to those associated with other methods that we believe to be plausible – those associated with the IPCW method (HR 0.76, 95% CI 0.36 – 1.62) and the per-protocol censoring method (HR 0.79, 95% CI 0.40 – 1.55). It is important to note that the "treatment group (ii)" analyses incorporated recensoring and we did not test a version without recensoring because, while pre-recurrence switching appeared likely to have been random, post-recurrence switching was likely to have been associated with prognostic characteristics.

Finally, it is worth reiterating why we believe the IPCW and per-protocol censoring analyses provide useful and plausible results. Rather than attempt to account for or adjust for post-disease recurrence use of imatinib, these methods simply attempt to correct for pre-recurrence switching onto imatinib. Implicitly this assumes that there is no need to adjust for post-recurrence use of imatinib, but given that such treatment is common in practice this may be reasonable. Sunitinib has become widely available in the UK since the imatinib trial was undertaken and this makes the issue of realistic post-recurrence treatment pathways more complex, but given the lack of data on post-recurrence treatments collected in the imatinib trial it appears reasonable to instead focus upon adjusting for the pre-recurrence switching. The IPCW analysis allows us to do this without making assumptions around the commonality of the pre- and post-recurrence imatinib treatment effect.

In addition, 17 of the 18 control group patients who became eligible to switch onto imatinib before disease recurrence did so, and because eligibility was based upon a calendar date it can be reasonably assumed that the switching that occurred was essentially random. This means that the IPCW can be run without the need for time-dependent covariates (other than time since

randomisation), which makes the method less prone to bias due to unmeasured confounding. It also means that we can use the entire control group when estimating stabilised weights, again helping to reduce the scope for bias, especially because the switching proportion (17.3%) is low. If the switching mechanism meant that the observed switching truly was random, then simple per-protocol adjustment approaches – either excluding switchers, or censoring them at the time of switch – would provide unbiased estimates of the treatment effect. Our results presented in Section 3 show that these methods gave similar results to the IPCW analysis for RFS and OS. Of the simple per-protocol approaches the censoring method may be preferred because it makes use of data from switchers up to the point of switch whereas the exclusion method does not use this data. However, the IPCW method may still be preferred because it allows us to adjust for any chance baseline differences between control group switchers and non-switchers, and to account for the time since randomisation at which switching occurred. We assess the baseline characteristics of switchers compared to non-switchers in Section 4.4.

4.2 Common treatment effect in pre-recurrence switchers

In Section 4.1 we hypothesise that the RPSFTM and IPE "treatment group (i)" analyses without recensoring are appropriate for our analysis of RFS, provided that the treatment effect received by control group patients who switch onto imatinib prior to disease recurrence can be shown to be similar to the treatment effect received by patients initially randomised to the experimental group. It is impossible to test this exactly, due to problems associated with time-dependent confounding. However, given that switchers received imatinib *before* disease recurrence, it might be reasonable to assume that they had a similar capacity to benefit, and had similar characteristics, to patients initially randomised to the experimental group. To investigate this further, we compared RFS in the experimental group to RFS in switchers from the point of switch. Because only 17 patients switched, this analysis is highly uncertain. We ran a Cox model comparing RFS in the experimental group to RFS from the point of switch in switching patients, and included baseline covariates for region, age, sex, time since resection, tumour size, whether the tumour was located in the stomach, performance status, resection margin, and mitoses, to adjust for any differences between switchers and experimental group patients. The resulting HR was 0.98 (95% CI 0.42 – 2.30), which suggests that the RFS experience of switchers (post-switch) was very similar to that experienced by patients randomised to the experimental group. This provides support for the hypothesis that the RPSFTM and IPE "treatment group (i)" analyses provide appropriate estimates of the adjusted RFS HR associated with imatinib.

4.3 Common treatment effect in post-recurrence switchers

In Section 4.1 we hypothesise that the RPSFTM and IPE "treatment group (ii)" analyses are appropriate for our analysis of OS, provided that the treatment effect received by control group patients who switch onto imatinib after disease recurrence can be shown to be similar to the treatment effect received by patients initially randomised to the experimental group. Again, it is impossible to test this exactly, due to problems associated with time-dependent confounding. Unlike the similar assumption regarding pre-recurrence switching investigated in Section 4.2, this assumption may not have face-validity if it is assumed that patients who only receive imatinib after disease recurrence have a lower capacity to benefit from it. However, to investigate this further we compared post-recurrence survival in control group patients that received post-recurrence imatinib to post-recurrence survival in control group patients that did not receive post-recurrence imatinib. We ran a Cox model and included baseline covariates for region, age, sex, time since resection, tumour size, whether the tumour was located in the stomach, performance status, resection margin, and mitoses, to adjust for any baseline differences between switchers and control group non-switchers. Ideally we would also include time-dependent covariates measured at the time of disease recurrence, but we did not have access to these. Instead, we included a covariate representing the RFS time for each patient which may be an important indicator of prognosis at the point of recurrence. This analysis involved an analysis of 36 control group patients who experienced disease recurrence and subsequently received imatinib, compared to 24 control group patients who experienced disease recurrence and did not subsequently receive imatinib. Only 17 deaths were observed and so the results of this analysis are highly uncertain. However, the resulting HR was 0.70 (95% CI 0.15 – 3.10), which is similar to (but actually numerically lower than) the point estimate for the adjusted OS HR estimated for experimental group imatinib according to the IPCW method. While the confidence interval is very large, this analysis suggests that there is not strong evidence against the assumption that post-recurrence switchers received a similar OS treatment effect to those who were initially randomised to the experimental group. This provides support for the hypothesis that the RPSFTM and IPE "treatment group (ii)" analyses provide appropriate estimates of the adjusted OS HR associated with imatinib.

4.4 Baseline characteristics of switchers

In Section 4.1 we hypothesised that the IPCW method provided an appropriate method to adjust for pre-disease recurrence in the imatinib trial, both for RFS and OS, primarily because the switching mechanism appeared to have been random. If this were truly the case, simple per-protocol censoring and exclusion methods could also appropriately be used to adjust for the switching. However, in a trial with a relatively small sample size there may have been chance differences

between patients who switched and patients who did not switch, particularly because only 17 patients switched. If this is the case, an IPCW analysis is likely to be more appropriate than a simple per-protocol analysis because it adjusts for any chance baseline differences between switchers and non-switchers, and also accounts for the time since randomisation that patients switched at. Table 3 summarises the baseline values of key baseline prognostic covariates measured in the imatinib trial.

Table 3: Summary of baseline values of key covariates by switcher group

Variable	All non-switchers (n=184)	Switchers (n=17)
Region		
West	21.5%	11.8%
Midwest	28.3%	41.2%
South	18.5%	5.9%
Northeast	19.0%	29.4%
Canada	13.0%	11.8%
Time since resection (days)	54.23	44.29
Tumour size		
Not marked	0.0%	0.0%
3cm – < 6cm	13.0%	0.0%
6cm – < 10cm	33.7%	23.5%
≥ 10 cm	53.3%	76.5%
Tumour located in stomach		
No	57.1%	64.7%
Yes	42.9%	35.3%
Sex		
Female	46.7%	41.2%
Male	53.3%	58.8%
Age (mean, years)	57.5	54.12
Performance status 0 at baseline		
No	28.8%	41.2%
Yes	71.2%	58.8%
Resection margin R0		
No	11.4%	11.8%
Yes	88.6%	88.2%
Mitosis category		
≤ 5	13.6%	23.5%
> 5 - ≤ 10	18.5%	35.3%
>10	64.1%	41.2%
Missing	3.8%	0.0%

Table 3 shows that there appear to have been some potentially important differences in baseline characteristics between switchers and non-switchers. Tumours appear to have generally been larger in switchers, and a larger proportion were located in the stomach. Fewer switchers had a baseline performance status category of zero, and switchers generally had a lower mitosis category. Potentially, then, there were important baseline differences between switchers and non-switchers.

This provides support for the hypothesis that an IPCW analysis should be preferred over simple per-protocol methods – which are likely to be prone to selection bias.

5. Conclusions

In this report we have conducted analyses to adjust for the switching observed in the imatinib trial. All switching adjustment methods are subject to important limitations, and their results are likely to be approximations of the truth. In some circumstances some methods are likely to be appropriate, whereas others are likely to be inappropriate. While no method may be perfect, they may provide more appropriate estimates of the treatment effect than that provided by the ITT analysis.

In this case study, we have demonstrated that the RPSFTM and IPE algorithm methods are only likely to be valid if applied in specific ways because of the trial characteristics and the data collected. Of these randomisation-based methods, for the analysis of RFS, we believe that the RPSFTM and IPE "treatment group (i)" approaches best fit the imatinib trial characteristics, given the data collected. Recensoring involves a loss of important information, and in the imatinib trial it is reasonable to assume that switching pre-recurrence was not associated with time-dependent covariates, and thus recensoring is not required. Using the RPSFTM "treatment group (i)" without recensoring analysis, we estimate an adjusted RFS HR associated with pre-recurrence treatment with imatinib of 0.52 (95% CI 0.32 – 0.85). The estimation procedure associated with this analysis performed better than that associated with the IPE "treatment group (i)" without recensoring analysis, which resulted in a HR of 0.55 (confidence intervals not calculated). These represent a numerical reduction in the HR compared to that estimated by the ITT RFS analysis (HR 0.60, 95% CI 0.42 – 0.89). These analyses are reliant upon the assumption that the treatment effect received by control group patients who switched onto imatinib before disease recurrence was similar to that received by patients initially randomised to the experimental group. Although this cannot be accessed with certainty, our analysis conducted in Section 4.2 suggests that this assumption is reasonable.

We believe that the IPCW analysis also provides an appropriate adjusted estimate of the RFS treatment effect, and because this method does not require a "common treatment effect" assumption, it may be preferable to the RPSFTM "treatment group (i)" analysis. The IPCW adjusted RFS HR is 0.50 (95% CI 0.32 – 0.78). The maximum weight estimated in the analysis was 1.47, indicating that no observations for any patients were allocated a particularly high weight, thus reducing the chance of error. The IPCW analysis is particularly well suited to this case study because pre-recurrence switching cannot have been associated with time-dependent prognostic characteristics, because eligibility to switch was dependent upon calendar time, and all but one patient who became eligible to switch did so. Given this, it might be hypothesised that a simple per-protocol censoring method may be sufficient to adjust for the observed pre-recurrence switching. However, due to chance differences in baseline characteristics in switchers compared to non-

switchers (as shown in Section 4.4), such a simplistic approach may remain prone to selection bias. Hence, the IPCW analysis – which also accounts for the time since randomisation at which switching occurred – is preferred.

For OS, we have again demonstrated that the RPSFTM and IPE algorithm methods are only likely to be valid if they are applied in specific ways. Of these randomisation-based methods, for the analysis of OS, we believe that the RPSFTM and IPE "treatment group (ii)" approaches best fit the imatinib trial characteristics, given the data collected. In this analysis recensoring is required, because control group patients who receive imatinib after disease recurrence are classified as switchers (although we do not adjust the survival times of these patients when estimating the adjusted treatment effect, because we only wish to adjust the treatment effect for pre-recurrence switching), and post-recurrence switching may have been associated with prognosis. This method is reliant upon the OS treatment effect received by patients who switched onto imatinib after disease recurrence being similar to the OS treatment effect received by patients initially randomised to the experimental group. Again, this assumption cannot be tested exactly, but our analysis presented in Section 4.3 suggests that there is not strong evidence against this assumption. Despite this, we acknowledge that this assumption might not be clinically plausible. For this OS analysis, it is also important to acknowledge that various post-study treatments may have been given, and these may have differed by randomised group. Given no further information on these treatments (aside from who received post-recurrence imatinib treatment in the control group) we have had to assume that similar post-study treatments were used in each group, such that no adjustments for these need to be made in order to estimate the OS treatment effect associated with imatinib compared to placebo. Using the RPSFTM "treatment group (ii)", we estimate an adjusted OS HR associated with pre-recurrence treatment with imatinib of 0.82 (95% CI 0.42 – 1.62). The estimation procedure associated with this analysis performed marginally better than that associated with the IPE "treatment group (ii)" analysis, which resulted in a HR of 0.81 (confidence intervals not calculated). These represent a numerical reduction in the HR compared to that estimated by the ITT OS analysis (HR 0.93, 95% CI 0.47 – 1.83).

As for the RFS analysis, we believe that the IPCW method also provides an appropriate adjusted estimate of the OS treatment effect, and because this method does not require a "common treatment effect" assumption, it may be preferable to the RPSFTM/IPE "treatment group (ii)" analyses. The IPCW adjusted OS HR is 0.76 (95% CI 0.36 – 1.62). The maximum weight estimated in the analysis was 1.63, indicating that no observations for any patients were allocated a particularly high weight, thus reducing the chance of error. Again, we believe that this analysis is preferable to a

simple per-protocol censoring analysis (which provided an HR of 0.79, 95% CI 0.40 – 1.55), although the similarity in the IPCW and per-protocol censoring estimates allows us to be more confident in our results.

References

[1] Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting survival time estimates to account for treatment switching in randomised controlled trials – an economic evaluation context: Methods, limitations and recommendations. Med Decis Making, January 21, 2014, doi:10.1177/0272989X13520192

[2] Morden JP, Lambert PC, Latimer NR, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Medical Research Methodology* 2011; 11.

[3] National Institute for Health and Clinical Excellence. Final appraisal determination: Everolimus for the second-line treatment of advanced renal cell carcinoma. 2010. London, NICE.

[4] GlaxoSmithKline UK. Pazopanib (Votrient(R)) for the first-line treatment of patients with advanced renal cell carcinoma (RCC): Addendum to GSK's submission to NICE. 2010.

[5] Robins JM, Tsiatis AA. Correcting for Noncompliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. *Communications in Statistics-Theory and Methods* 1991; 20(8):2609-2631.

[6] Mark SD, Robins JM. A Method for the Analysis of Randomized Trials with Compliance Information - An Application to the Multiple Risk Factor Intervention Trial. *Controlled Clinical Trials* 1993; 14(2):79-97.

[7] White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. *Statistics in Medicine* 1999; 18(19):2617-2634.

[8] Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Statistics in Medicine* 2002; 21(17):2449-2463.

[9] White IR. Estimating treatment effects in randomized trials with treatment switching. *Statistics in Medicine* 2006; 25(9):1619-1622.

[10] Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; 56(3):779-788.

[11] Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* 2001; 96(454):440-448.

[12] Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part I: Structural nested models and marginal structural models to test and estimate treatment arm effects. *Statistics in Medicine* 2004; 23(13):1991-2003.

[13] Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. *Statistics in Medicine* 2004; 23(13):2005-2022.

[14] Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias. *American Journal of Epidemiology* 2011; 173(5):569-577.

[15] Stata statistical software intercooled, Version 11.0, Texas, USA: 2009.

[16] Fewell Z, Hernan MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal* 2004; 4(4):402-420.

[17] White IR, Walker S, Babiker AG. strbee: Randomization-based efficacy estimator. *The Stata Journal* 2002; 2(2):140-150.

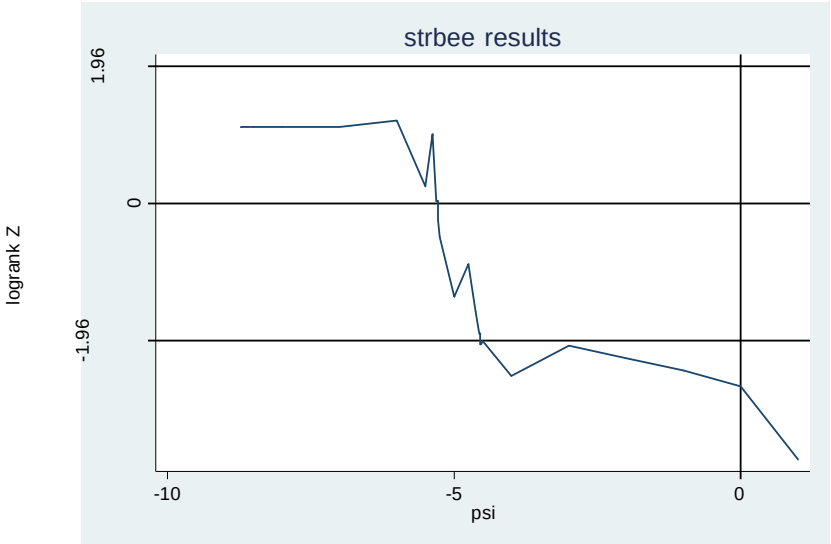
[18] Robins JM, Greenland S. Adjusting for Differential Rates of Prophylaxis Therapy for Pcp in High-Dose Versus Low-Dose Azt Treatment Arms in An Aids Randomized Trial. Journal of the American Statistical Association 1994; 89(427):737-749.

Appendix 1. Log-rank test results for alternative values of psi

Psi ( ) equals the –ln(acceleration factor). The RPSFTM estimate of psi is taken as the value for which the logrank Z statistic is equal to 0. The upper and lower confidence intervals are taken as the value of psi for which the logrank Z statistic is equal to ±1.96.

RFS

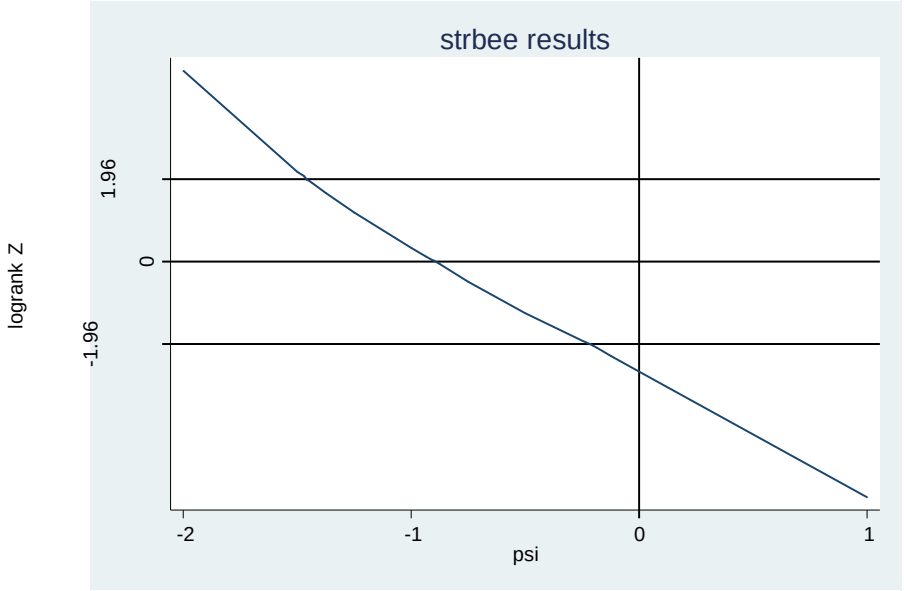
Analysis 4: RPSFTM "on treatment" with recensoring



Analysis 5: RPSFTM "on treatment" without recensoring
No convergence.

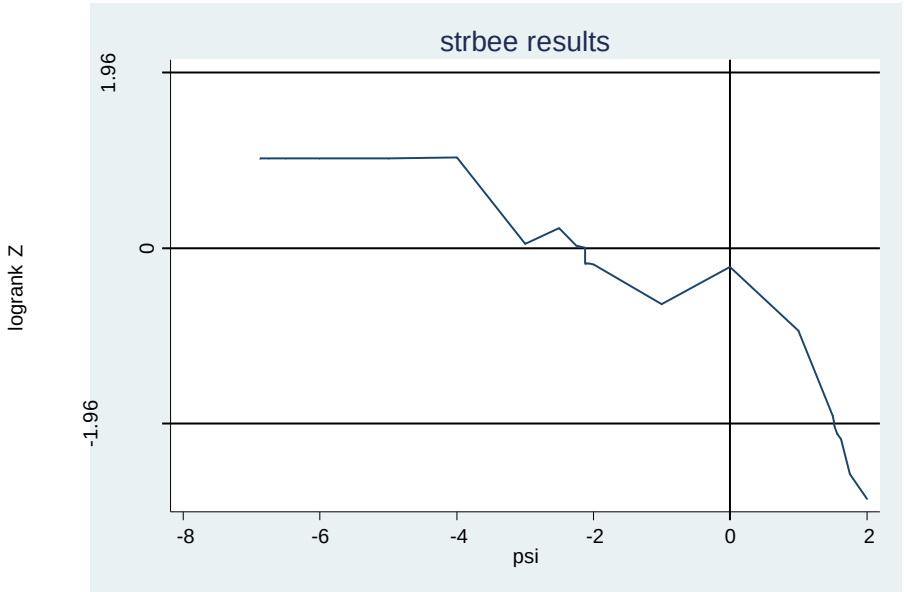
Analysis 6: RPSFTM "treatment group (i)" with recensoring

Analysis 7: RPSFTM “treatment group (i)” without recensoring

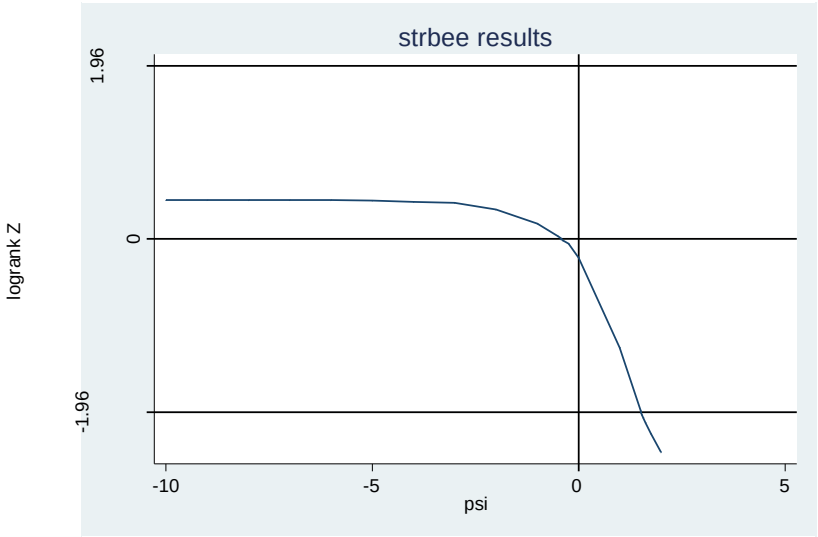


OS

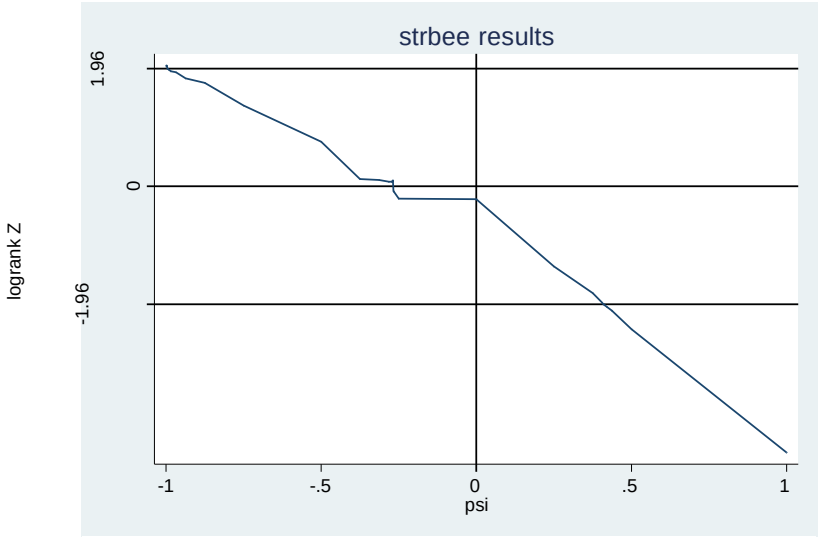
Analysis 16: RPSFTM “on treatment” with recensoring



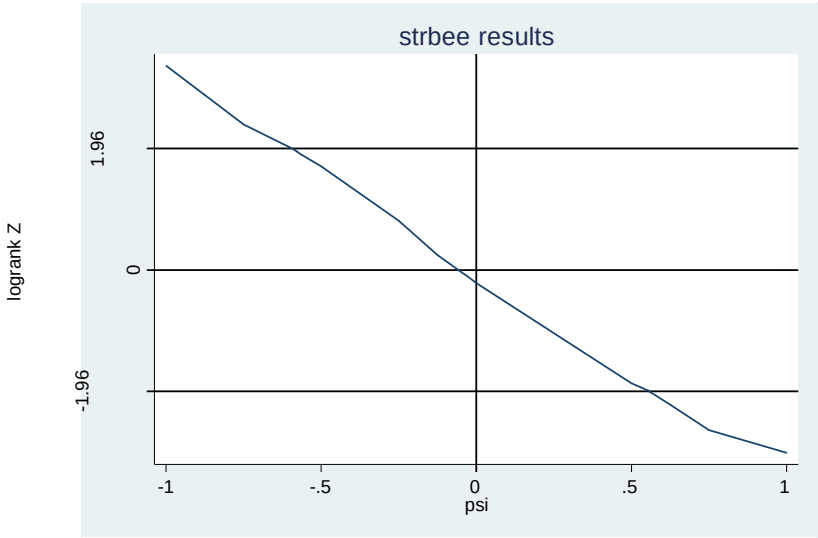
Analysis 17: RPSFTM “on treatment” without recensoring



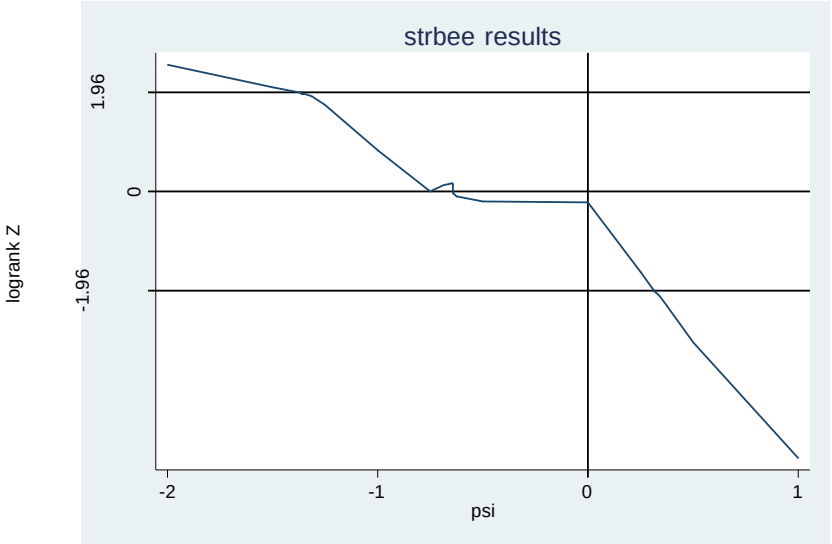
Analysis 18: RPSFTM “treatment group (i)” with recensoring



Analysis 19: RPSFTM “treatment group (i)” without recensoring



Analysis 20: RPSFTM “treatment group (ii)” with recensoring

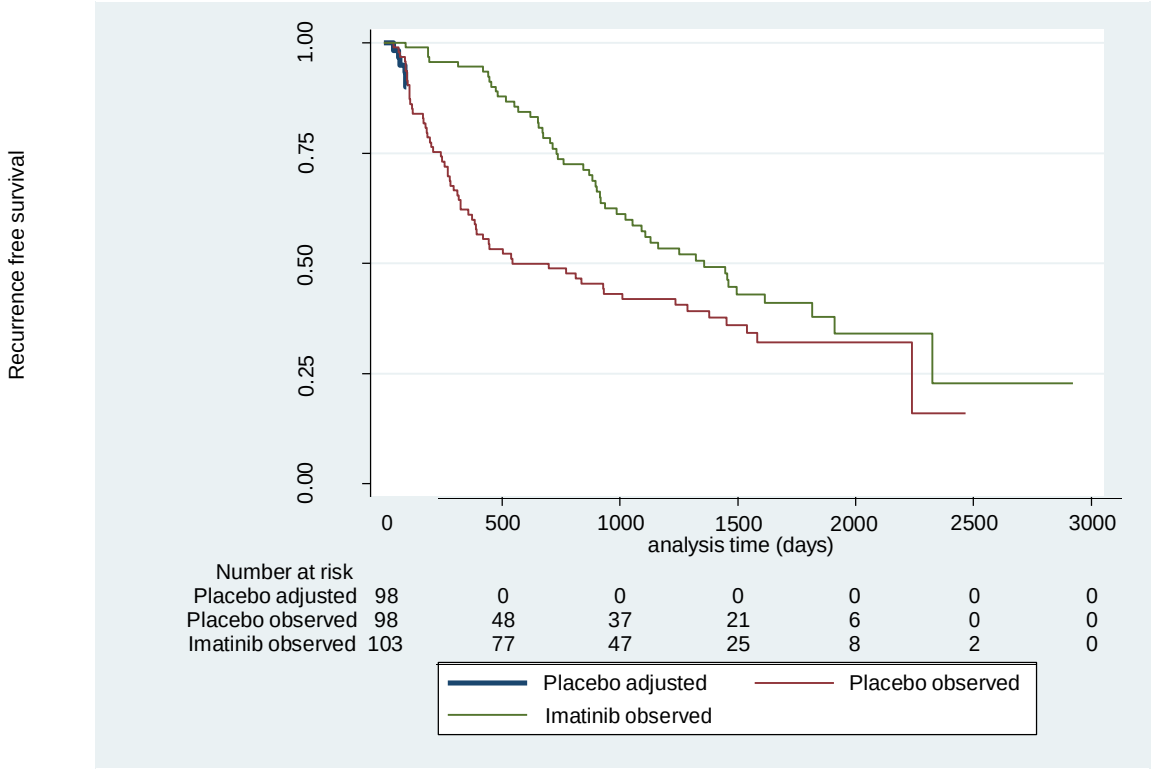


Appendix 2: Kaplan-Meier curves from RPSFTM and IPE analyses

In each graph the "placebo adjusted" curve represents the Kaplan-Meier curve for the control group adjusted to account for the pre-recurrence treatment switching.

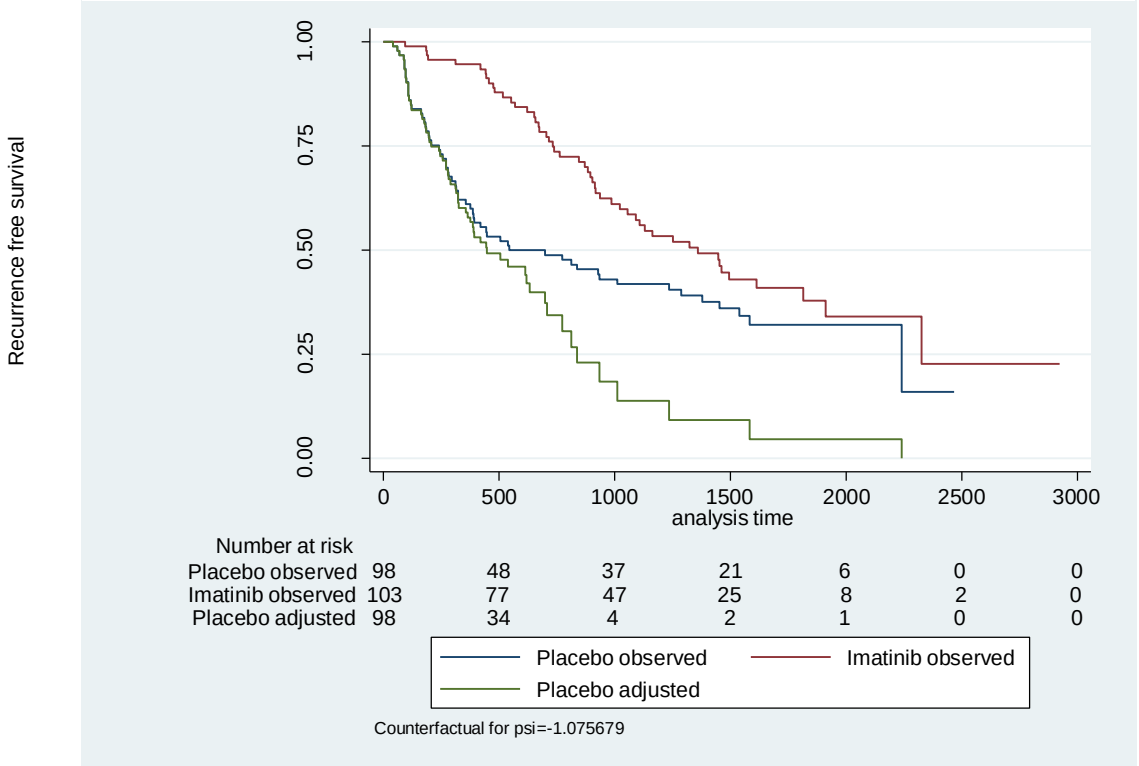
RFS

Analysis 4: RPSFTM "on treatment" with recensoring

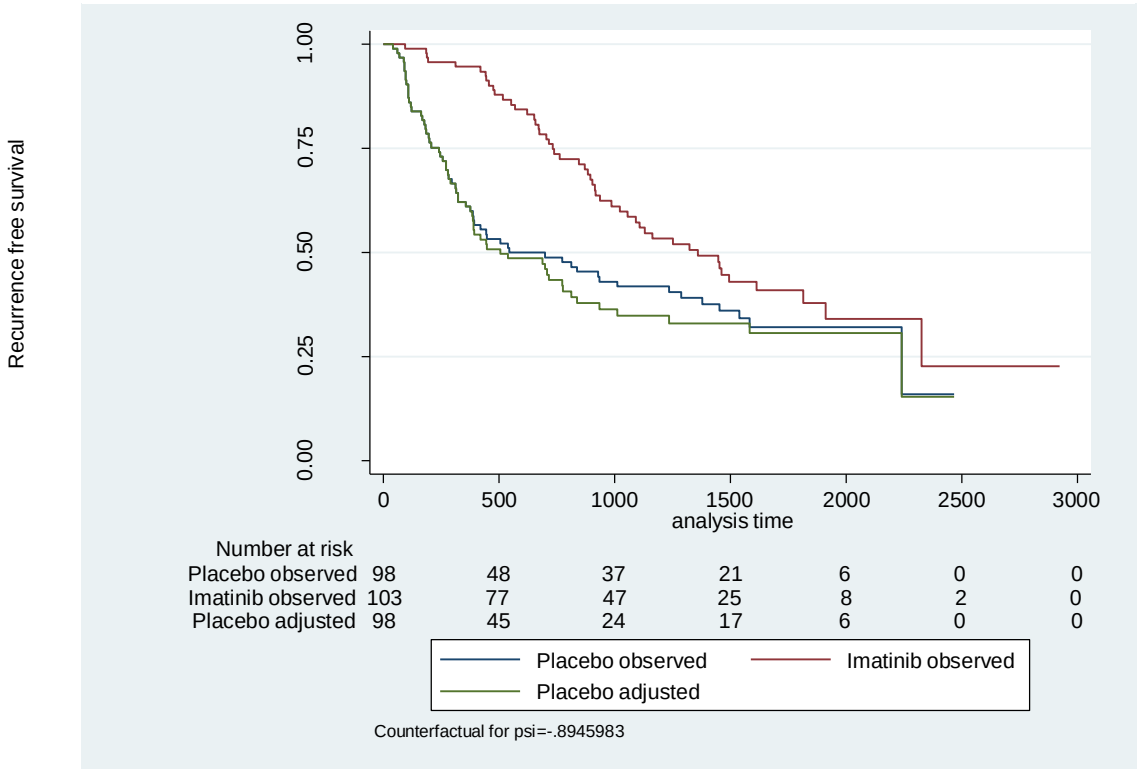


Analysis 5: RPSFTM "on treatment" without recensoring
No convergence.

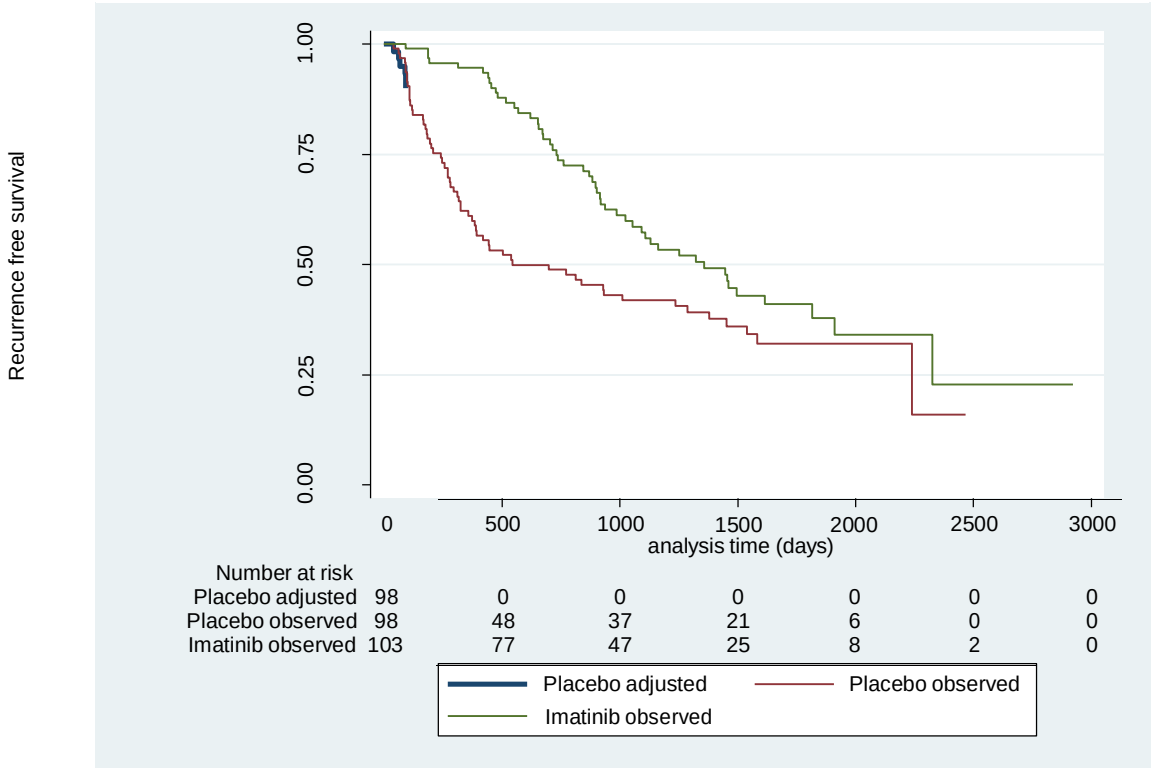
Analysis 6: RPSFTM “treatment group (i)” with recensoring



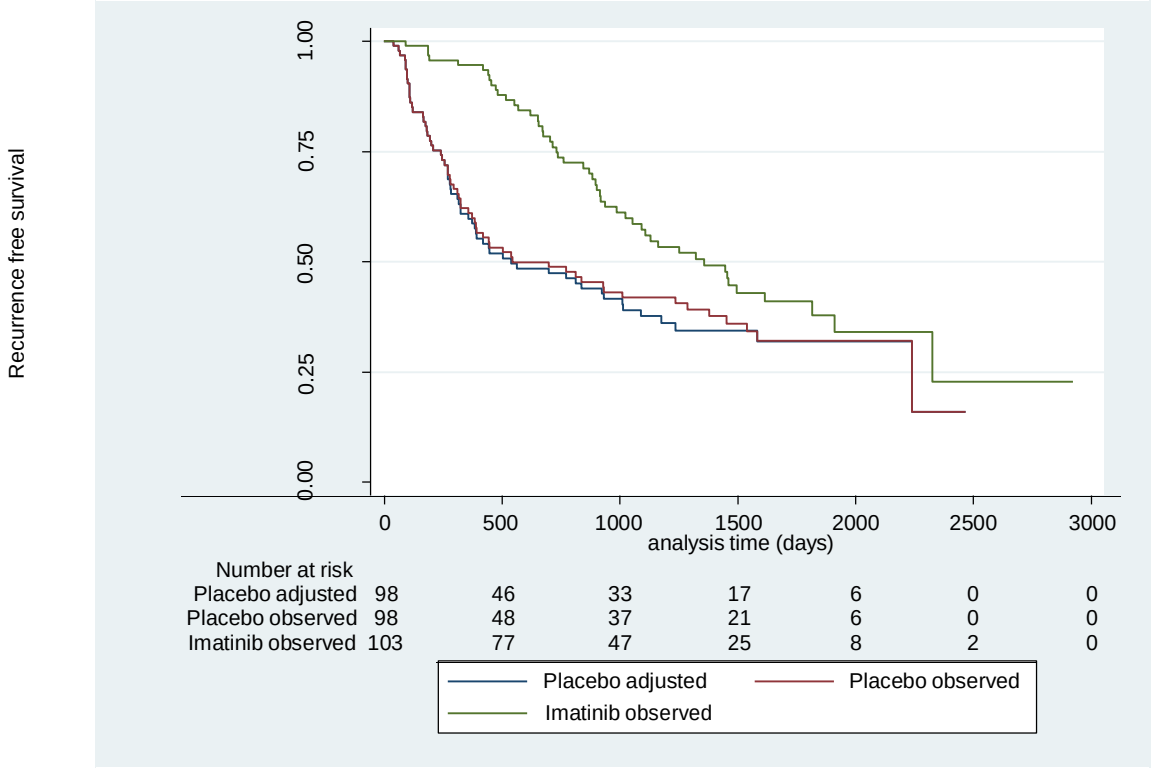
Analysis 7: RPSFTM “treatment group (i)” without recensoring



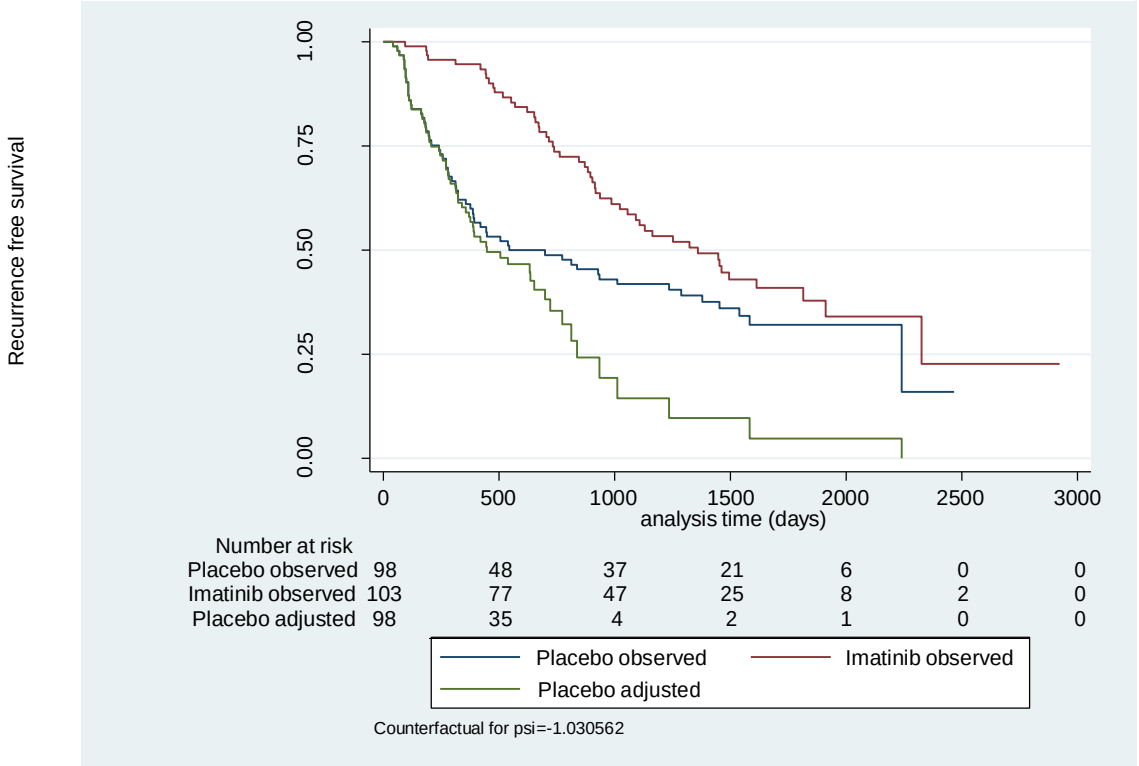
Analysis 8: IPE “on treatment” with recensoring



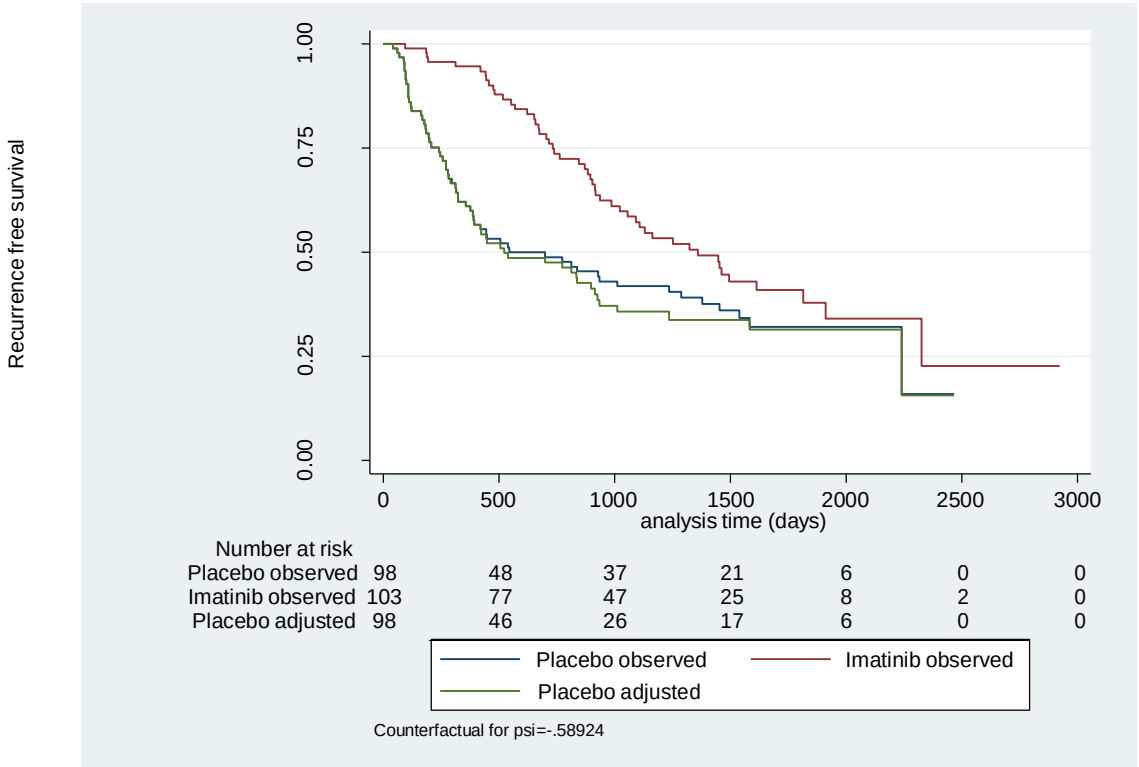
Analysis 9: IPE “on treatment” without recensoring



Analysis 10: IPE “treatment group (i)” with recensoring

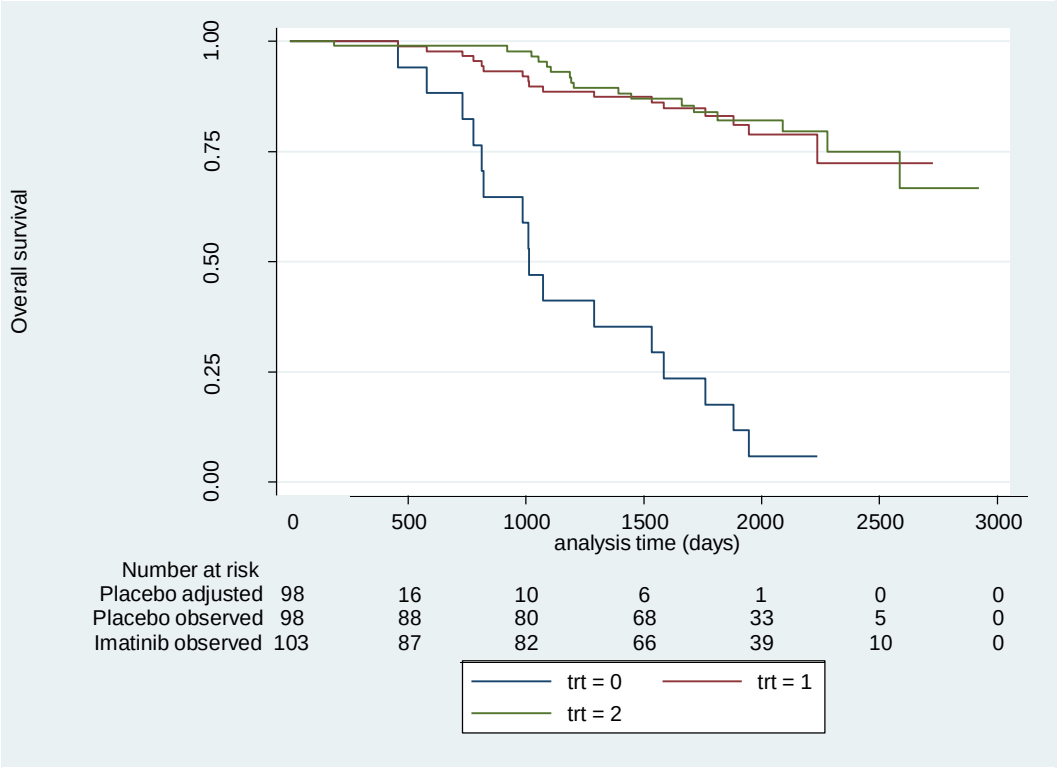


Analysis 11: IPE “treatment group (i)” without recensoring

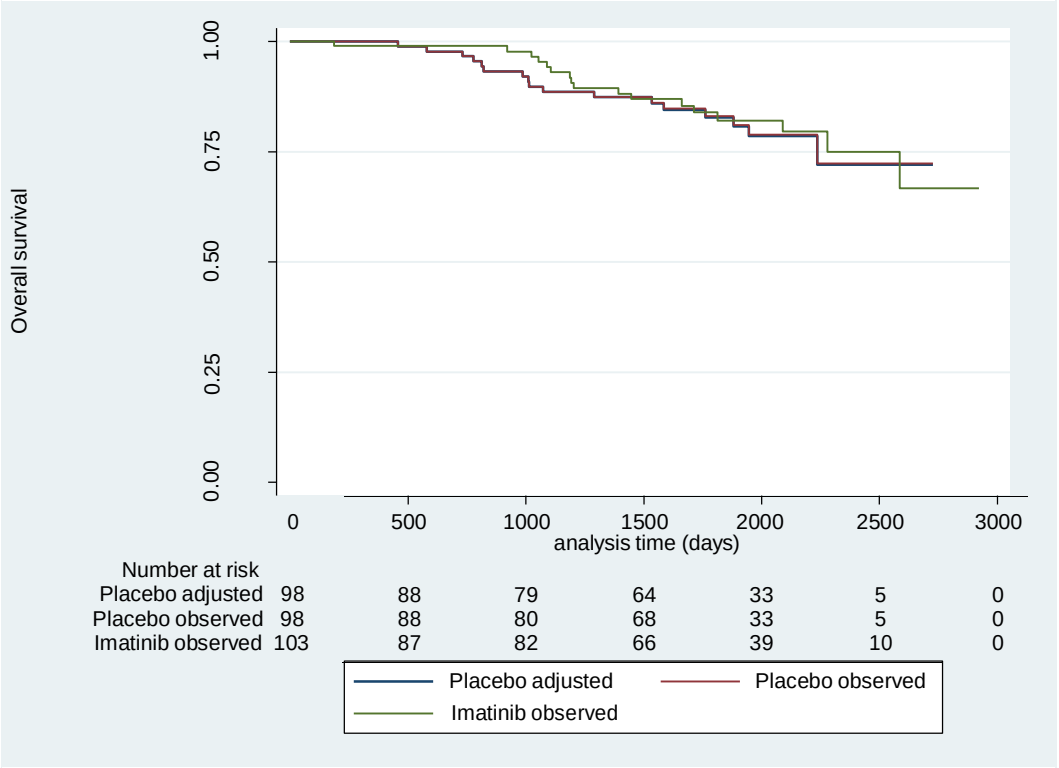


OS

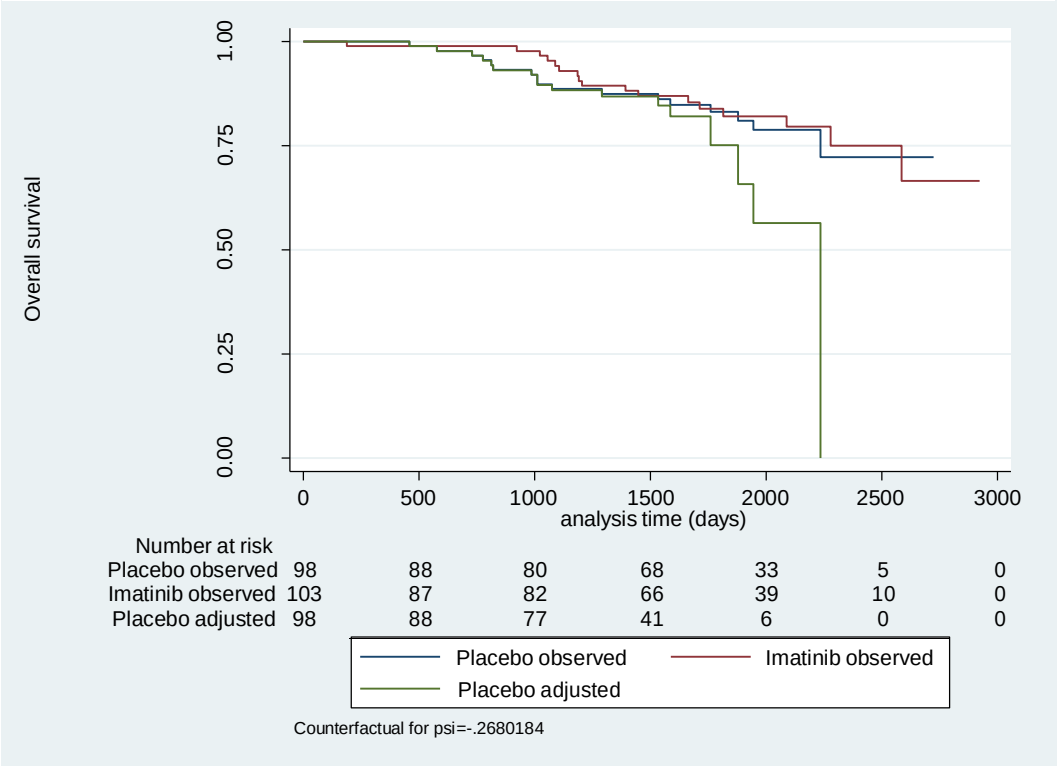
Analysis 16: RPSFTM “on treatment” with recensoring



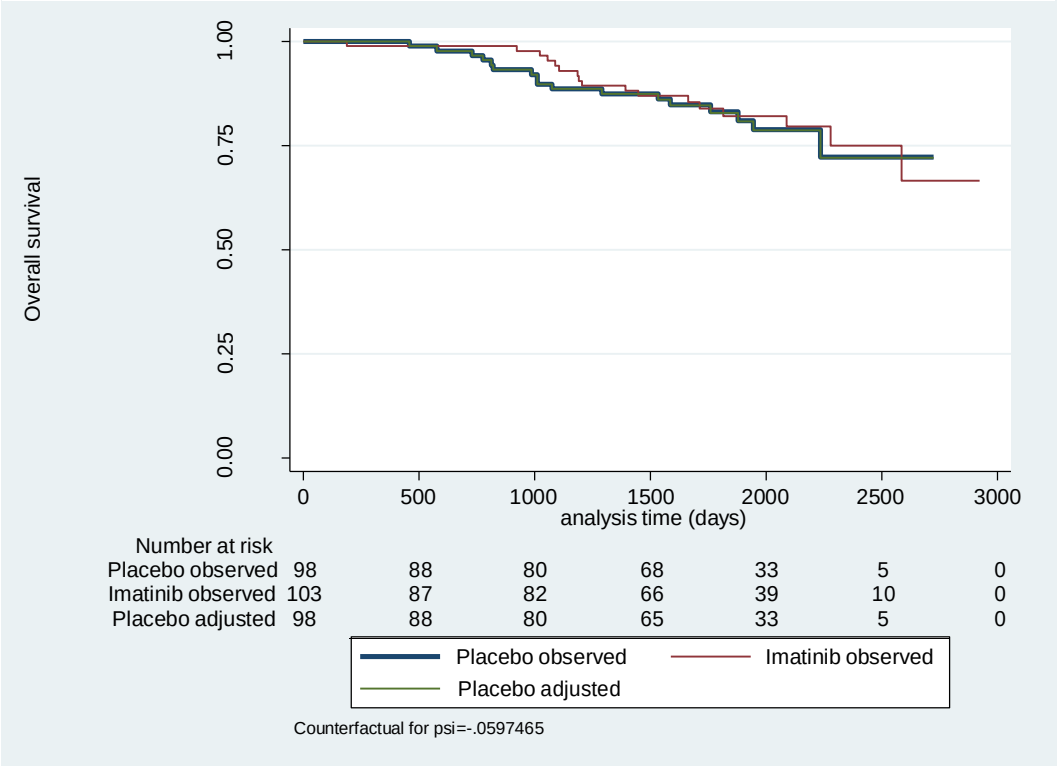
Analysis 17: RPSFTM “on treatment” without recensoring



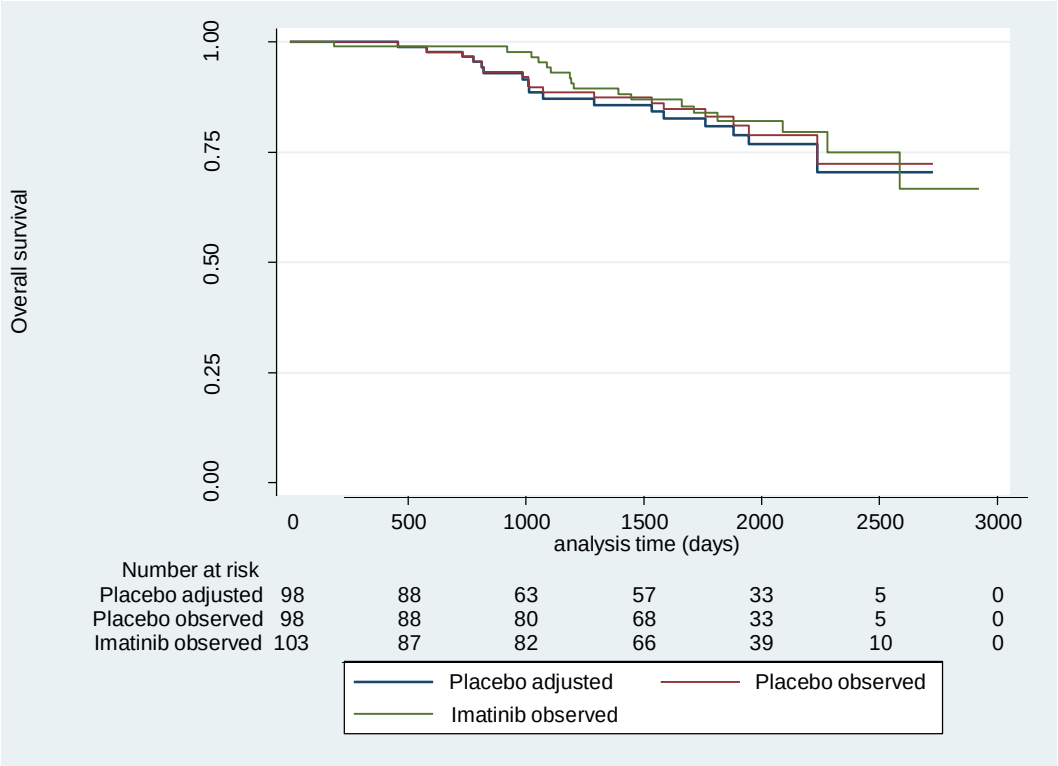
Analysis 18: RPSFTM “treatment group (i)” with recensoring



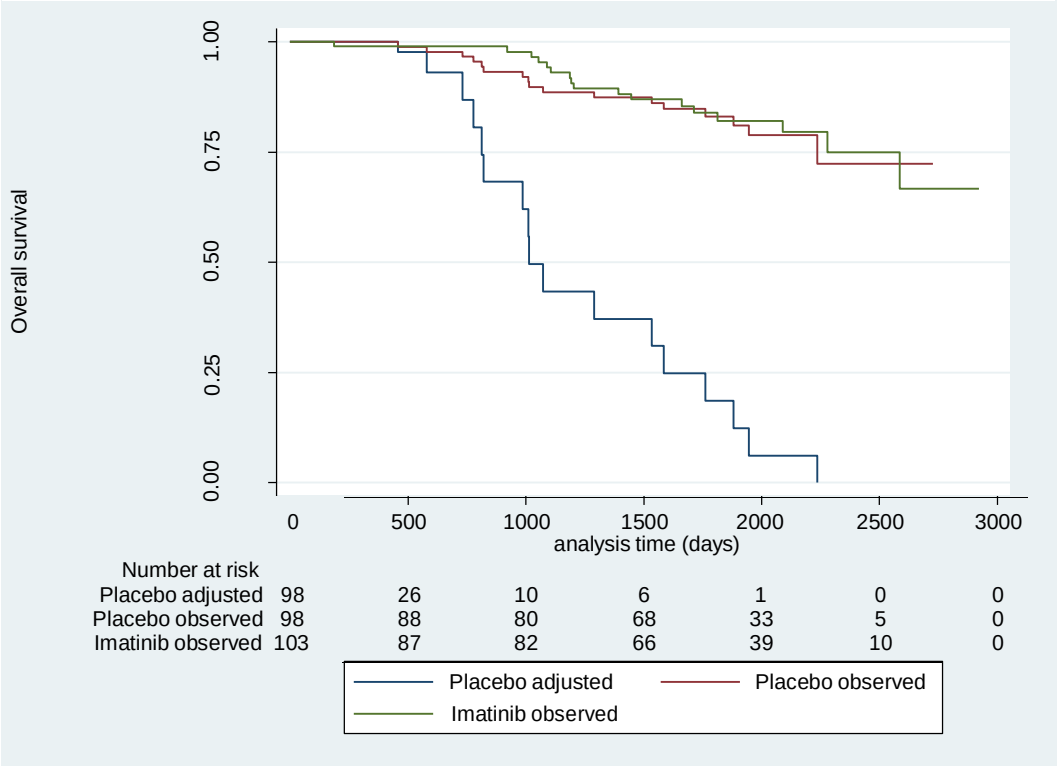
Analysis 19: RPSFTM “treatment group (i)” without recensoring



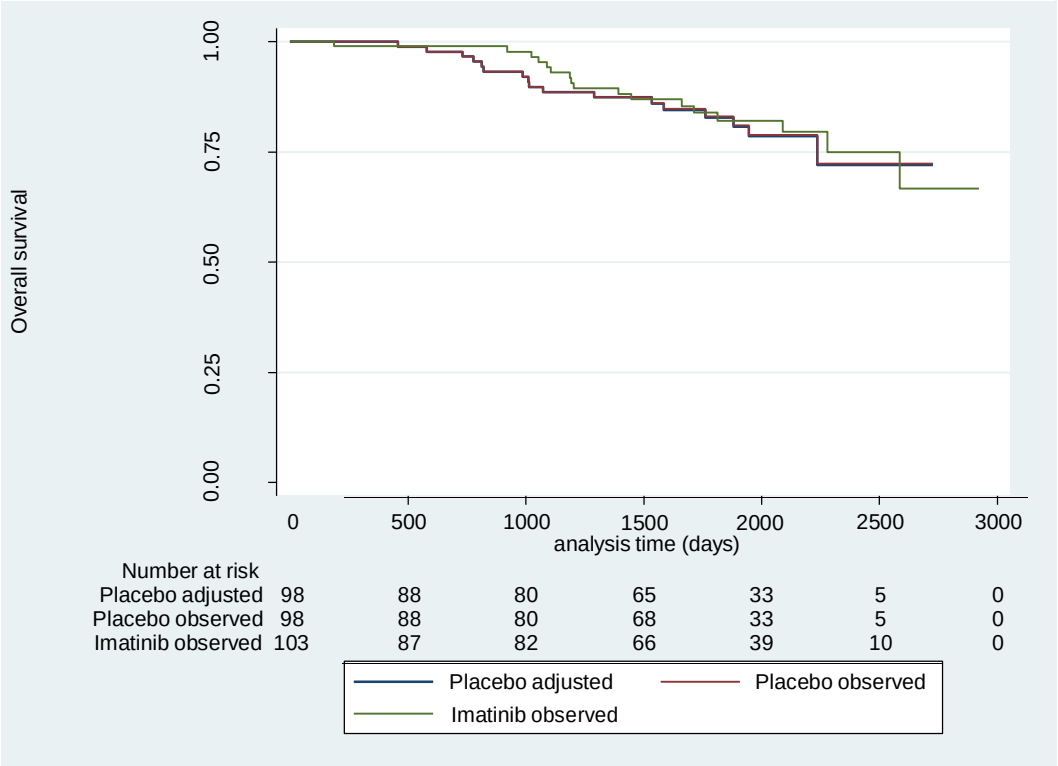
Analysis 20: RPSFTM “treatment group (ii)” with recensoring



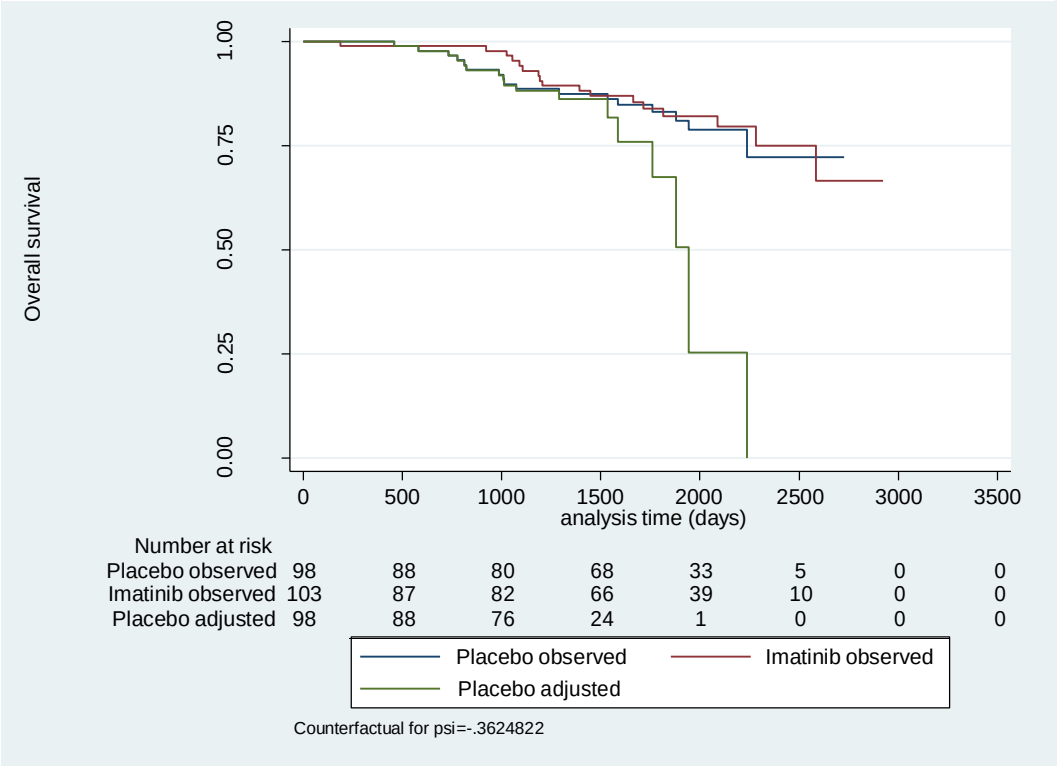
Analysis 21: IPE “on treatment” with recensoring



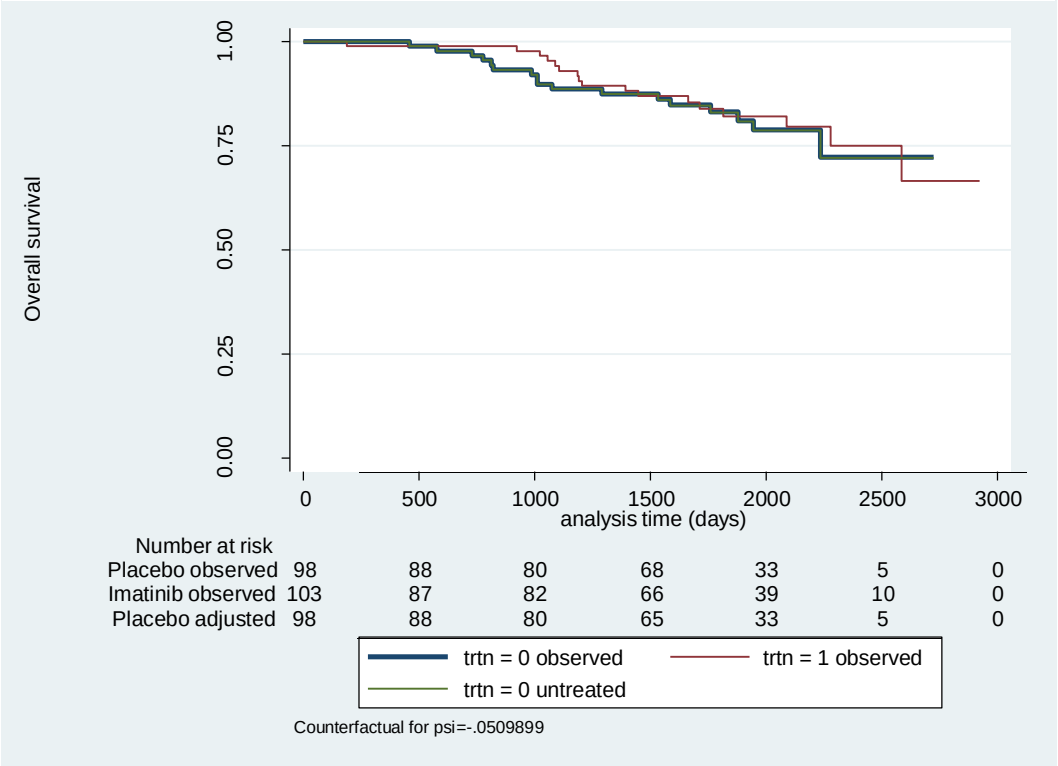
Analysis 22: IPE “on treatment” without recensoring



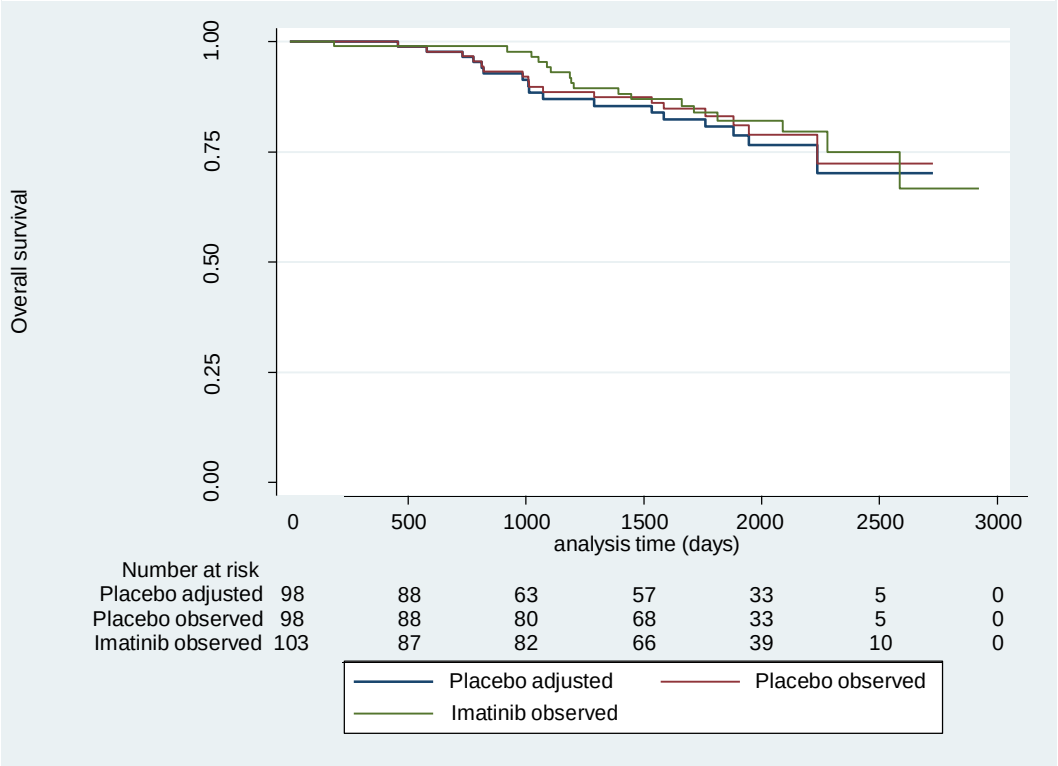
Analysis 23: IPE “treatment group (i)” with recensoring



Analysis 24: IPE “treatment group (i)” without recensoring



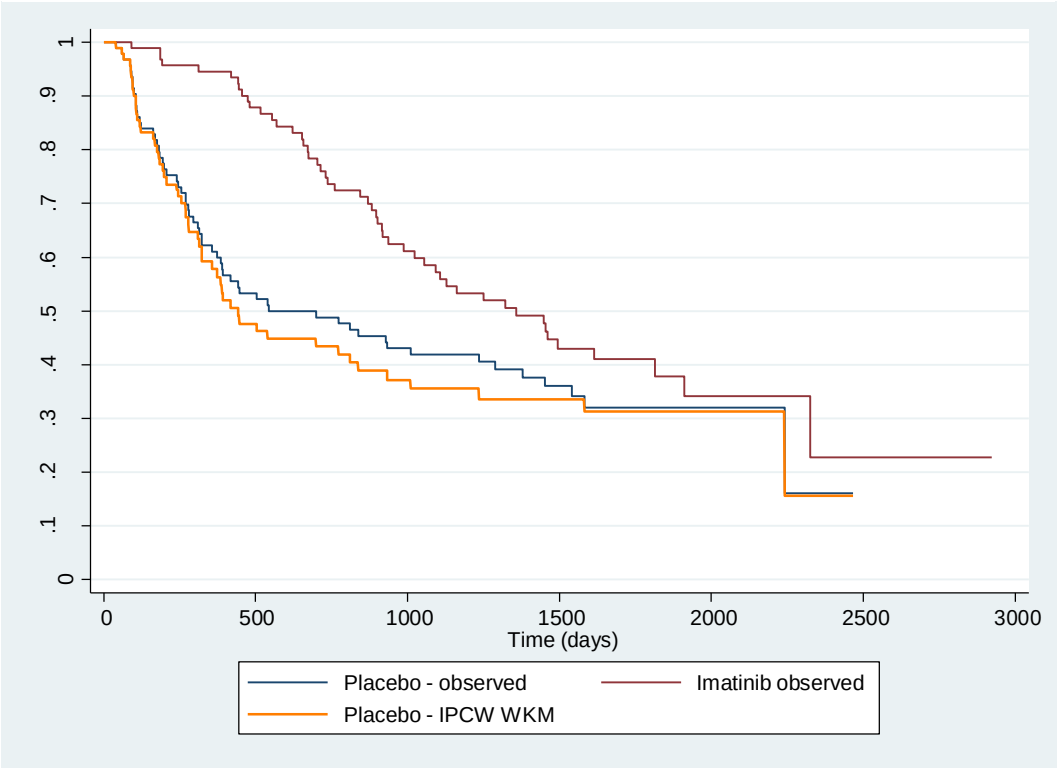
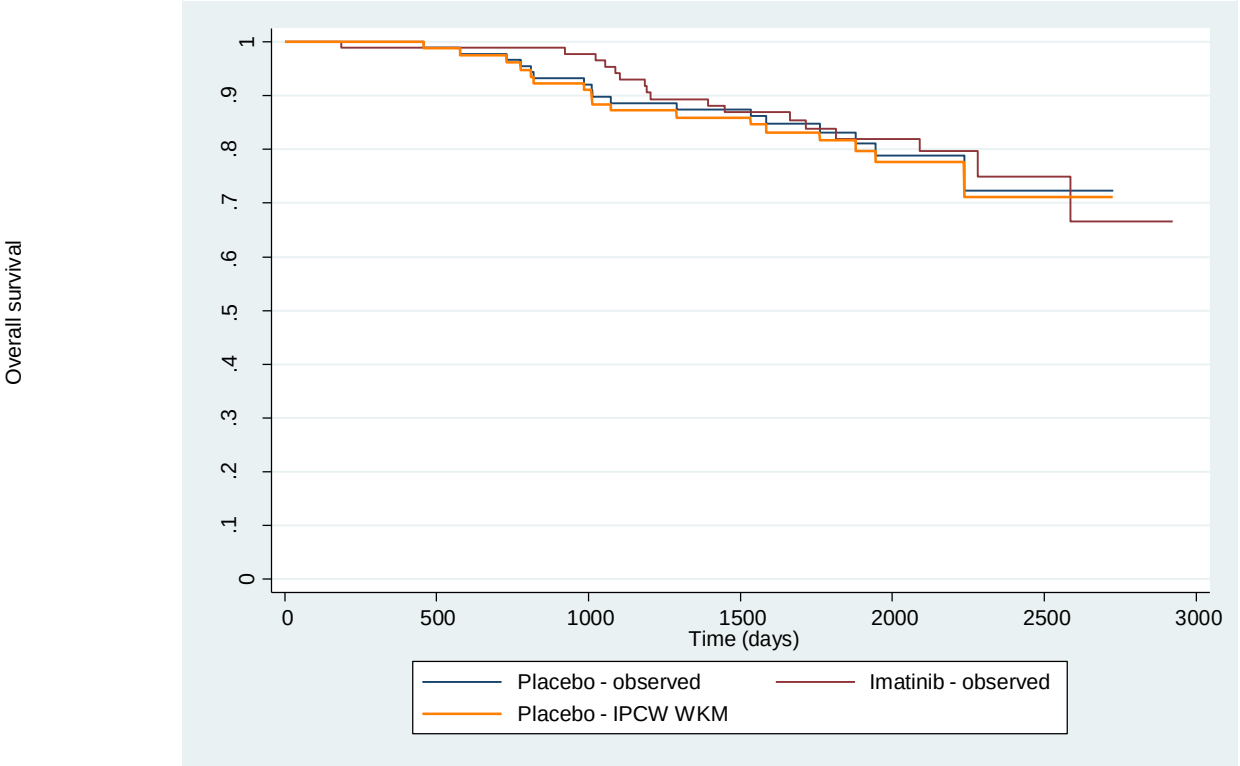
Analysis 25: IPE “treatment group (ii)” with recensoring



Appendix 3: Weighted Kaplan-Meier curves from IPCW analyses

It is worthy of note that the IPCW adjusted estimate of the HR incorporates stabilised weights to account for potential time-dependent selection bias caused by censoring switching patients, *and* also incorporates baseline covariates in the regression that estimates the adjusted HR. The IPCW weighted Kaplan-Meier (WKM) curves incorporate the stabilised weights, but do not additionally adjust for baseline covariates. Hence they may be perceived differences between the treatment effect illustrated by the WKM and the treatment effect calculated using the IPCW HR.

Analysis 12: IPCW RFS analysis



Analysis 26: IPCW OS analysis